Mining protein-protein interaction networks for the analysis of disease

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Dr. rer. nat.
im Fach Biophysik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Dipl.-Inform. Martin Schaefer

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Stefan Hecht, PhD

Gutachter:
1. Prof. Dr. Hanspeter Herzel
2. Dr. Miguel Andrade
3. Prof. Dr. Matthias Selbach

eingereicht am: 25.09.2012
Tag der mündlichen Prüfung: 01.03.2013

Acknowledgements

I would like to thank a number of people who helped me in various ways while I was conducting my PhD studies and writing this thesis. First of all I would like to thank Miguel Andrade for supervising my thesis, numerous discussions and excellent support while I was working in his research group. I learned a lot during these years and I am deeply thankful for his inspiration and motivation.

I would also like to thank the past and current members of Miguel's group for making the years in the lab a joyful time. Among those, special thanks goes to Matt for reading and correcting my thesis.

I was involved in a very close and fruitful collaboration with the group of Prof. Erich Wanker. I would like to thank him and his group for sharing their thoughts and data.

I also would like to thank my family, especially my parents and my grandmother, for the support and education they gave me.

Finally, I would like to thank my friends both inside and outside of the Max-Delbrück-Center for helping me to find the right balance between research and personal life.

Abstract

Interactions between proteins regulate signaling, gene expression and many other cellular functions. Therefore, characterizing the entire human interactome is a key effort in current proteomics research. The existing knowledge of protein-protein interactions (PPIs) is stored in a number of databases. However, PPIs have properties that make their interpretation difficult and that are not adequately represented in a unified way in these databases. On the one hand, the experimental reliability of the techniques used to detect PPIs can have widely different quality with some methods being associated with high error rates. Another problem of PPI detection methods is that many interactions are measured under artificial conditions (for example, yeast cells are transfected with human genes in yeast two-hybrid assays) or even if detected in a physiological context, this information is missing from the common PPI databases.

We implement a resource that integrates human PPI data from the major expert-curated PPI databases. To address the high uncertainty associated with experimentally detected PPIs, we develop a scoring scheme that has been optimized both computationally and by human experts to reflect the amount and quality of evidence for a given PPI. To deal with the problem of missing context, we develop a method that assigns information to PPIs inferred from various attributes of the interacting proteins: gene expression, functional and disease annotations, and inferred pathways. We demonstrate that context annotation helps to detect interactions of higher experimental reliability and how context-filtered networks are enriched in bona fide pathways and disease proteins. We use these context-specific networks to identify PPIs that likely play a role in disease.

Finally, we use the integrated human PPI network for the study of the wild type function of polyglutamine (polyQ) stretches. Expansions of these stretches have been observed in the proteins of a large number of patients with different neurodegenerative diseases such as Huntington's and several Ataxias. Protein aggregation, which is a key feature of most of these diseases, is thought to be triggered by these expanded polyQ sequences in disease-related proteins. However, polyQ tracts are a normal feature of many human proteins, suggesting that they have an important cellular function. To clarify the potential function of polyQ repeats in biological systems, we study the characteristics of polyQ-containing proteins in the human PPI network. We complement the network analysis studying the repeats at nucleotide, protein and organism level. Together, our observations suggest that polyQ tracts in proteins stabilize protein interactions, likely through structural changes whereby the polyQ sequence extends a neighboring coiled-coil region to facilitate its interaction with a coiled-coil region in another protein.

Zusammenfassung

Interaktionen zwischen Proteinen regulieren Signalwege, Genexpression and viele andere zelluläre Funktionen. Die Charakterisierung der Gesamtheit menschlicher Proteininteraktionen gehört daher zu den wichtigsten Zielen der Proteomik. Das existierende Wissen über Proteininteraktionen wird in Datenbanken gespeichert. Allerdings haben Proteininteraktionen Eigenschaften, die ihre Interpretation und konsistente Repräsentation in Datenbanken erschweren. Zum einen besitzen Methoden zur Detektion von Proteininteraktionen stark variierende experimentelle Verlässlichkeit. Einige dieser Methoden sind mit sehr hohen Fehlerraten assoziiert. Andererseits werden viele Proteininteraktionen unter artifiziellen Bedingungen gemessen (beispielsweise werden beim Yeast-Two-Hybrid-Verfahren Interaktionen von menschlichen Proteinen in Hefezellen beobachtet) und selbst wenn sie unter natürlichen Bedingungen gemessen werden, fehlt eine Beschreibung des physiologischen Kontexts in den gängigen Interaktionsdatenbanken.

Wir implementieren eine Anwendung, die menschliche Proteininteraktionsdaten aus den wichtigsten, von Experten gepflegten Datenbanken integriert. Um die hohen Fehlerraten von experimentell detektierten Proteininteraktionen zu adressieren, entwickeln wir eine Funktion, die sowohl computergestützt als auch von Experten dahingehend optimiert wird, Menge und Qualität der Evidenz einer Proteininteraktion zu bewerten. Um das Problem der fehlenden Kontextinformationen zu beheben, entwickeln wir eine Methode, die Interaktionsannotationen von verschiedenen Attributen der interagierenden Proteine ableitet. Dazu berücksichtigen wir gewebespezifische Expression, Funktion und Krankheitsrelevanz der Proteine sowie vorhergesagte Signalwege, an denen die Proteine beteiligt sind. Wir zeigen, dass eine spezifischere Annotation einer Proteininteraktion mit höherer experimenteller Verlässlichkeit einhergeht und dass Netzwerke, die spezifisch sind für bestimmte Kontext-Typen, angereichert sind in kanonischen Signalwegen und krankheitsrelevanten Proteinen. Wir benutzen die kontextspezifischen Netzwerke, um Proteininteraktionen zu identifizieren, die vermutlich eine Rolle in Krankheiten spielen.

Schließlich verwenden wir das integrierte humane Netzwerk interagierender Proteine für die Untersuchung der Wildtyp-Funktion von Polyglutaminketten. Expansionen dieser Ketten wurde in Patienten mit verschiedenen neurodegenerativen Erkrankungen (wie zum Beispiel Chorea Huntington und mehreren Ataxien) beobachtet. Es wird angenommen, dass Proteinaggregation, ein Hauptmerkmal der meisten dieser Krankheiten, durch die Verlängerung der Polyglutaminketten in krankheitsrelevanten Proteinen ausgelöst wird. Allerdings sind Polyglutaminketten normaler Bestandteil vieler menschlicher Proteine, was suggeriert, dass diese Ketten eine wichtige zelluläre Funktion haben. Um Hinweise auf eine solche Funktion in biologischen Systemen zu sammeln, untersuchen wir die Charakteristika von Proteinen mit Po-

lyglutaminketten in Interaktionsnetzwerken. Diese Analyse ergänzen wir durch eine Untersuchung der Sequenzwiederholungen auf Nukleotid-, Protein- und Organismen-Ebene. Zusammengenommen legen unsere Beobachtungen nahe, dass Polyglutaminketten Interaktionen zwischen Proteinen stabilisieren. Wahrscheinlich erfolgt diese Stabilisierung durch Veränderungen in der Proteinstruktur, wobei die Polyglutaminsequenz eine benachbarte Coiled-Coil-Region erweitert, um die Interaktion mit einer Coiled-Coil-Region in einem anderen Protein zu ermöglichen.

Contents

1 Introduction

1.1 Motivation

Almost all cellular processes involve proteins in one way or another. Proteins bind to each other and form complex networks of protein-protein interactions (PPI) to achieve these functions. Accordingly, many research efforts focus on the discovery PPIs. New experimental methods are constantly being developed to measure PPIs and the amount of available PPI data is steadily increasing. Currently, expert-curated databases contain tens of thousands of human PPIs.

At the same time, many experimental PPI detection techniques have high error rates resulting in a large number of wrongly reported interactions. Additionally, most PPIs are detected under conditions that are to a certain degree artificial, thus hindering the interpretation of the generated data. For example, in yeast two-hybrid (Y2H), which is currently the most frequently applied assay for the detection of direct interactions between human proteins, the candidate proteins are expressed in yeast to monitor a potential binding event. For non-yeast proteins, this leads to many reported interactions that, under physiological conditions, might never occur or occur in only a very limited number of cell types.

While for systems level analyses of the PPI network, where the focus is on global properties of the interactome (e.g., for the detection of hub proteins or frequent associations of protein domains in interacting protein pairs) the effects of false positive interactions might be negligible, for studies focusing on small numbers of interactions (e.g., for the detection of causative PPIs in disease) the large number of technical and biological false positives will hinder the generation of reliable hypotheses. In general, the development of methods and tools to identify and specifically analyze PPIs of higher technical reliability and biological relevance lags behind the rate by which new PPI data is generated.

In this work, we first present a method to identify interactions supported by more reliable experimental evidence and then assign biological context information to PPIs. This allows us to generate context-specific, high-confidence PPI networks. We will illustrate the usefulness of this approach by demonstrating how the resulting networks allow one

to highlight pathway information or disease-relevant interactions.

1.2 Overview

Chapter 2 starts with an overview of the biology of proteins and their interactions. It introduces basic concepts related to the generation and analysis of PPI networks both from an experimental and a computational point of view. Advantages and disadvantages of the most commonly applied experimental techniques to measure PPIs are discussed. The major projects that collect and provide experimentally generated PPI data are introduced. The basics of graph theory and common PPI network analysis tasks are explained. The chapter closes with a discussion of the relevance of PPIs for disease.

Chapter 3 addresses the problem of high uncertainty in PPI networks by defining a scoring scheme for published PPIs that reflects the amount of evidence supporting each interaction. This scoring method is implemented as a web tool providing a resource that automatically retrieves the newest interactions from expert-curated databases, integrates them into a single database and scores them according to our method. We will also discuss problems associated with the integration of multiple PPI networks: both research interests and technical limitations bias most networks that are currently being reported. We will quantify the extent to which this affects large integrated networks such as ours.

After developing a method to select reliable interactions, in Chapter 4 we present a strategy to detect interactions that are relevant to a specific problem by annotating interactions with context information (e.g., from expression profiles or functional information of the participating proteins). We illustrate how the annotation with context information in combination with network algorithms is able to select interactions of high biological relevance. We then use our method to detect interactions that play a role in the crosstalk between influenza proteins with innate immune response and phosphorylation-dependent PPIs related to Alzheimer's disease.

In Chapter 5 we address an open biological question that has been discussed for many years in the scientific community: whether there is a wild-type function of polyglutamine (polyQ). By providing evidence that polyQ-containing proteins underlie evolutionary selection and illustrating that polyQ proteins have distinctive features in the PPI networks of several species, we develop the hypothesis that polyQ has a specific function in mediating PPIs.

2 Protein-protein interactions

2.1 Cellular function emerges from the interaction between proteins

Proteins are among the most important molecules in living organisms and achieve many cellular functions. As indicated by the protein database UniProt (Apweiler et al., 2011), more than 20,000 human proteins with diverse biological roles are known: approximately 10% are involved in binding of the DNA to control gene expression, 7% are receptors, which sense and report the presence of ligands, and 3% are kinases, which catalyze the transfer of a phosphate group (usually from the coenzyme ATP) to a substrate.

Proteins are composed of combinations of 20 different amino acids each of which has different biochemical properties (such as charge, hydrophobicity and polarity), which determine the protein's structure and function. Proteins usually do not act alone but carry out their function in cooperation with other proteins. When this cooperation involves a specific physical binding event between the participating proteins it is termed a PPI. Here, we use the term PPI to distinguish from a mere functional relation where both proteins are involved in the same cellular function but not necessarily physically interact or from a genetic interaction where mutations in two or more genes lead to the same phenotype. In the following, interaction (or interactome for the sum of all interactions in an organism) refers to PPI unless otherwise stated.

Many cellular functions are achieved by the complex interplay between proteins. To name a few examples:

- PPIs mediate the transduction of signals from the outside to the inside of a cell where a physiological response happens. The external signals are sensed by receptor proteins, which often control the activity of intracellular kinases. The kinases in turn propagate the signal and modify the activity of target proteins. These cascades of PPI events are termed signaling pathways.

- The transcriptional machinery, which controls gene expression, consists of large protein complexes, which dynamically assemble in the nucleus of the cell.

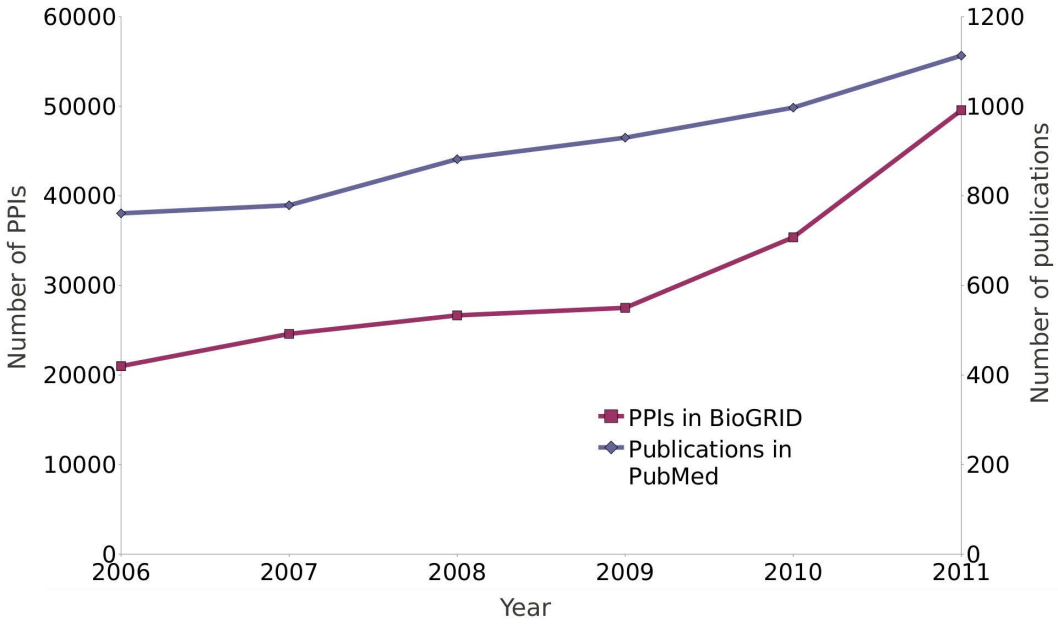

Figure 2.1: Increase of human PPIs stored in the manually curated PPI databases BioGRID and search results in PubMed for 'protein-protein interaction' per year.

- Structural components of the cell rely on PPIs: Actin assembles into microfilaments, which form part of the cytoskeleton. The motor protein myosin realizes muscle fiber contraction by moving along such actin filaments.

All of these functions require the highly specific recognition of protein interaction partners. Furthermore, these protein binding events underlie delicate control mechanisms and are dependent on the type and state of the cell in which they occur.

Accordingly, only a small fraction of all possible combinations of protein pairs interact. Estimates of the size of the human interactome range from around 130,000 to 260,000 interactions (Hart et al., 2006; Venkatesan et al., 2009). Over the last years, the amount of PPI data increased steadily (see Figure 2.1). Even though the cumulative amount of reported interactions approaches the estimated lower limit of the human interactome size, due to the high error rates of the experimental techniques, the majority of PPIs likely remain to be discovered.

Knowledge of PPIs is crucial for the understanding of protein and cellular function. Therefore, completing the characterization of the human interactome is a key effort in proteomics. Many experimental methods exist to detect interactions and algorithms have been developed to interpret the generated PPI network data.

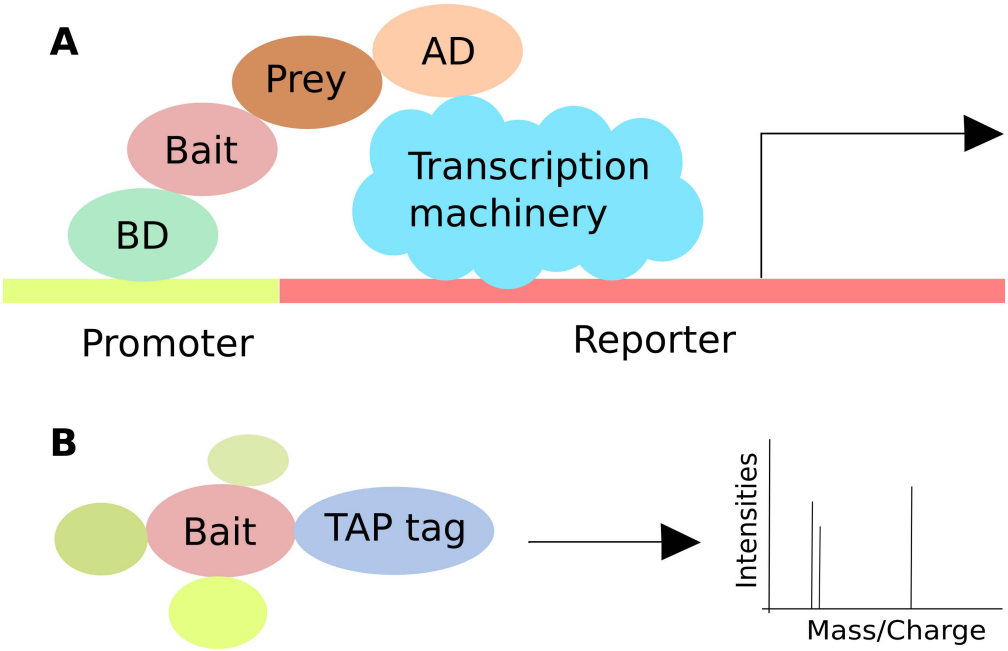

Figure 2.2: Basic principles of two commonly applied methods for the detection of PPIs. (A) Y2H detects direct physical binding events between a bait and a prey protein while (B) TAP/MS reports the composition of complexes in which the bait protein is found.

2.2 Detection methods

PPIs can be measured by many experimental methodologies, which have hugely different experimental set-ups and degrees of confidence. They either detect a direct physical binding event between two proteins (binary methods) or isolate a complex whose protein composition is subsequently determined (co-complex methods). The protein (or group of proteins) whose interaction partners are to be determined is termed the bait protein. The potential interaction partners are called prey proteins.

The most commonly applied method for detecting binary PPIs is Y2H, which was first applied in the late 1980s (Fields and Song, 1989). It is a genetic high-throughput method that is based on the fact that some eukaryotic transcription factors, such as the yeast transcriptional activator GAL4, are composed of two subunits, which need to assemble for the regulation of transcription. GAL4 consists of a DNA-binding domain (BD) and an activation domain (AD). To detect the interaction between a bait and a prey protein, the bait is fused to the BD and the prey to the AD. The fusion constructs are expressed in the same yeast cell, which contains a reporter gene that is activated

by a GAL4 responsive promoter. Neither of the two GAL4 subunits alone is sufficient to drive the expression of the reporter gene. Only if the bait and prey protein interact and, in doing so, bring the BD and the AD into close proximity a functional activator is formed and the expression of the reporter gene driven (Figure 2.2A).

Y2H can be automated up to genome-wide scale by expressing bait and prey proteins in haploid yeast strains of opposite mating type. The yeast strains are combined and the resulting diploid yeast clones coexpress the combination of bait and prey proteins. As a result, large amounts of proteins can be screened for interactions at the same time at relatively low costs. Another advantage of Y2H is its sensitivity towards transient interactions since the reporter gene expression significantly amplifies the signal (Estojak et al., 1995). A general drawback of Y2H is that it detects interactions outside their normal environment (e.g., in case of non-yeast proteins or even yeast proteins if they are cytoplasmic). As a consequence, proteins might not undergo correct post-translational modifications (PTMs) and folding. Also, multiple studies estimated high false-positive rates for Y2H (Hart et al., 2006; Mrowka et al., 2001; Von Mering et al., 2002). Usually, error rates are deduced from the overlap between the PPI networks generated in different studies. Venkatesan et al. (2009), in contrast, explained the low overlap between different studies with a low sensitivity (by missing many interactions) rather than with a lack of specificity.

In several seminal studies Y2H has been applied to generate draft PPI maps of entire organisms: *Saccharomyces cerevisiae* (Ito et al., 2001; Uetz et al., 2000), *Drosophila melanogaster* (Formstecher et al., 2005) and human (Rual et al., 2005; Stelzl et al., 2005). Additionally, disease networks (Kaltenbach et al., 2007; Lim et al., 2006) and functional subnetworks (Bandyopadhyay et al., 2010; Colland et al., 2004; Lehner and Sanderson, 2004) have been elucidated using large-scale Y2H approaches. As of today, around 31% of all interactions in human PPI databases were measured with Y2H (Schaefer et al., 2012a) placing this method number one among the most frequently applied PPI detection techniques in humans.

Another method for the detection of direct physical PPIs is luminescence-based mammalian interactome mapping (LUMIER) (Barrios-Rodiles et al., 2005): a bait protein is fused to luciferase, an enzyme that catalyzes a light-emitting reaction. Another protein is fused to a FLAG-tag, which allows one to capture the tagged protein. An interaction can be detected by a luminescence signal. An obvious advantage of LUMIER is that it allows the study of mammalian PPIs in mammalian cell types.

LUMIER was used to identify the components of the TGF-β (Barrios-Rodiles et al., 2005) and the Wnt (Miller et al., 2009) signaling pathways. The method was also used

to validate PPIs previously measured with Y2H (Braun et al., 2008).

In contrast to the previously described binary approaches, co-complex methods require an initial purification step in which the protein complex is isolated. The complex composition is then characterized, e.g., using mass spectrometry (MS) where the protein identity of complex members is revealed by measuring the characteristic mass-to-charge ratio of charged peptides. For complex purification, the bait protein is either directly targeted with an antibody (co-immunoprecipitation) or fused to a tag, which is then captured rather than the bait protein itself. Tandem affinity purification (TAP) (Puig et al., 2001; Rigaut et al., 1999) is a more stringent purification strategy in which two sequential affinity tags are fused to the bait protein (Figure 2.2B). The two successive purification steps decrease the number of non-specific binding partners but may also remove weak or transient interaction partners of the bait protein (Von Mering et al., 2002).

Several large scale studies identified protein complexes in yeast (Gavin et al., 2002, 2006; Krogan et al., 2006) and *Escherichia Coli* (Butland et al., 2005) using TAP/MS. In human cell lines, TAP/MS was used to identify complexes involved in specific pathways (Bouwmeester et al., 2004; Major et al., 2007) and in the transcription machinery (Jeronimo et al., 2007).

Another MS-based method that aims to overcome the problem of missing weak interactions while removing contaminants is quantitative proteomics. Here, the relative abundance of interaction partners of a tagged protein in comparison to a control experiment (e.g., the tag alone or RNAi knockdown of the target protein) is determined. A recently developed variant of quantitative proteomics is stable isotope labeling by amino acids in cell culture (SILAC) (Mann, 2006) in which two populations of cells (one expressing the tagged bait protein and the other expressing the control) are grown: one in light and one with heavy amino acid medium (most commonly different isotopes of arginine or lysine). This differential labeling leads to a mass shift of the proteins of the cell that can be detected by MS and allows to determine the required abundances (Vermeulen and Selbach, 2009). The stringent purification in TAP/MS and the quantitative approach correcting for unspecific binding in SILAC are commonly believed to increase the specificity of PPI detection importantly.

Large scale interaction maps among kinases (Oppermann et al., 2009) and proteins involved in the cell cycle (Olsen et al., 2010) have been generated using SILAC in human cell lines. Differences in the relative abundance ratios were exploited to distinguish between stable and transient interactions (Wang and Huang, 2008).

In addition to experimental methods, computational methods have been developed

that predict protein interactions based, for example, on orthology, protein sequence, domain composition, co-expression and functional annotations. Sometimes, combinations of these features are applied to predict novel interactions or to estimate the reliability of experimentally measured PPIs (Brown and Jurisica, 2005; Jensen et al., 2009). In several species, similarity in the Gene Ontology (GO) term description of the protein pair and coexpression over many conditions or tissues have been shown to be good predictors of interactions (Lu et al., 2005; Maetschke et al., 2012; Qi et al., 2006). Sequence and the presence of potentially interacting domains generally perform worse (Ta and Holm, 2009; Yu et al., 2010). Recently, physical docking methods were shown to produce good results in the prediction of PPIs (Wass et al., 2011). This method, however, is limited by the computational complexity and the large number of proteins for which the tertiary structure has not yet been resolved.

Different decisions on the study design introduce technical and selection biases, which lead to an enrichment of certain protein classes and pathways in the measured PPI networks (Futschik et al., 2007; Von Mering et al., 2002):

1. TAP/MS has a high sensitivity towards abundant proteins. As a consequence, interactions between highly abundant proteins are more easily detected (Björklund et al., 2008; Ivanic et al., 2009; Von Mering et al., 2002).

2. Due to physiochemical constraints, different experimental methods preferentially detect interactions within subsets of proteins with certain properties. For example, Y2H tends to detect interactions between protein pairs located in the nucleus (Jensen and Bork, 2008) and TAP/MS reports with a higher frequency interactions involving small proteins under 15 kDa (Gavin et al., 2002).

3. Some proteins are more frequently studied than others. This selection bias is particularly strong for literature-curated (for example, Peri et al. (2003) preferentially consider disease-related genes for their expert-curated PPI database HPRD) and integrated networks (in Chapter 3 we will demonstrate that proteins are studied with largely different frequencies and how this affects integrated PPI databases).

2.3 Curation and availability

Experimentally detected PPIs are collected in several publicly available databases that are curated by experts and make the PPI's supporting evidence easily accessible. Usually, these databases provide meta-data such as the study in which the interaction has been described and which techniques have been applied to measure the interaction. They

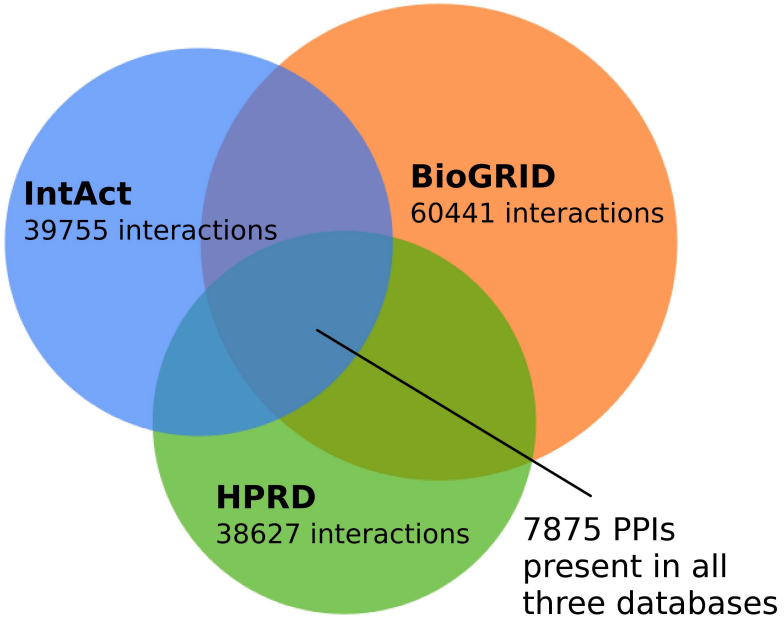

Figure 2.3: Agreement between the three largest publicly available expert-curated PPI databases HPRD, BioGRID and IntAct. Only human PPIs are considered.

implement different mechanisms to query and display the data. These databases include BioGRID (Stark et al., 2011), HPRD (Keshava Prasad et al., 2009), IntAct (Aranda et al., 2010) and MINT (Ceol et al., 2010).

As it has been previously noted, many of the interactions are unique to a certain database (Futschik et al., 2007; Lopes et al., 2011) and few interactions are listed in all of them. Figure 2.3 illustrates the current overlap for the three largest expert-curated databases. Accordingly, several databases integrate the data from multiple expert-curated resources: UniHI (Chaurasia et al., 2007) and iRefWeb (Turner et al., 2010) are comprehensive PPI resources integrating the major public PPI databases. The popular functional interaction resource STRING (Szklarczyk et al., 2011) additionally incorporates interactions computationally predicted by various methods.

To assess the quality of PPI data stored in public databases, Cusick et al. (2008) re-evaluated randomly selected PPIs from several manually curated resources. They found that both in human and yeast more than a third of the probed interactions were wrongly annotated, with the most frequent errors being wrongly assigned species, incorrectly reported protein identity and absence of adequate experimental settings. On top of experimental noise, these wrongly curated interactions add another layer of error. Despite the high reported error rates, expert-curated PPI databases are often used as

gold-standard datasets for various purposes such as estimating performance parameters of a screen (Bandyopadhyay et al., 2010) or training of classifiers for the prediction of PPIs (Geisler-Lee et al., 2007).

Von Mering et al. (2002) reported a higher accuracy for PPIs supported by multiple studies. Similarly, Cusick et al. (2008) found fewer wrongly annotated interactions among PPIs reported in several databases. Both observations open the possibility for evidence-based selection of likely true PPIs from the entire set of reported interactions. However, assigning scores that estimate the experimental reliability based on the cumulative evidence supporting the interaction remains challenging (Braun et al., 2008). Accordingly, only few manually curated databases associate reliability scores with PPIs. An exception is the database MINT, which scores interactions detected in small scale experiments higher than those from high-throughput experimental methods, considers the number of studies in which the interaction has been found and rates conservation of the interaction between homologous proteins (Ceol et al., 2010). IntAct recently announced that it will also release an evidence-based confidence score soon (Kerrien et al., 2012).

An important step for allowing comparability between different studies, work groups and databases has been the standardization attempts of the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI). They released the Proteomics Standards Initiative Molecular Interaction (PSI-MI) standard, which defines the representation of molecular interaction data and, by defining a controlled vocabulary, their annotation with information on the experiments conducted to measure the interaction (Hermjakob et al., 2004). Additionally, a PPI exchange standard has been developed: Proteomics Standards Initiative Common Query Interface (PSICQUIC) (Aranda et al., 2011). PSICQUIC specifies a protocol (accessible by the web services SOAP and REST) that standardizes access to several of the major PPI databases and enables automated data retrieval from multiple sources.

Due to the large number of interactions available in PPI databases, tools are required to display and layout the network data. A popular graph viewer for PPI networks is Cytoscape (Shannon et al., 2003). It allows one to render even large networks of thousands of interactions. It implements various layout algorithms, offers a basic network analysis utility and can be extended by a large number of plugins implementing all kinds of functionality ranging from interfaces to the aforementioned PPI databases to high-level network algorithms (examples will be given in the next section).

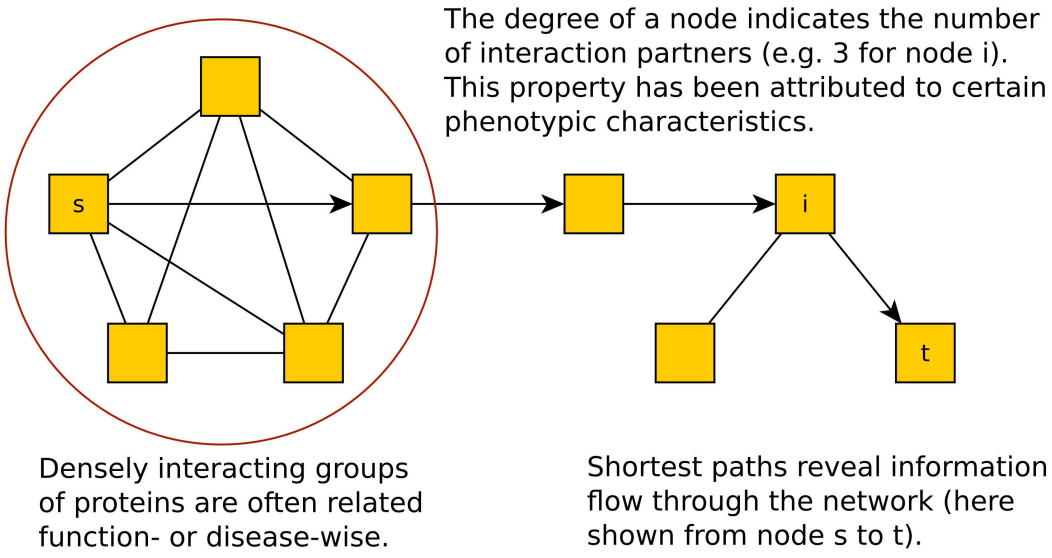

Figure 2.4: Topological properties used to characterize and identify essential components of PPI networks.

2.4 Network analysis

Various research questions related to the function of single or groups of interacting proteins can be addressed with the help of PPI networks. Therefore, physical interactions between proteins in a network are modeled in the formal framework of graph theory. A graph $G = \{V, E\}$ consists of a set of nodes V and a set of edges E that connect the nodes. Topological measures allow the characterization of the properties of the graph and network algorithms allow the identification of important components of the graph (Figure 2.4). When modeling PPIs, the nodes of a graph represent proteins and edges indicate interactions between proteins.

In graph theory directions can be assigned to edges. With respect to PPI networks, directed edges are commonly used to model signal flow between proteins.

Several studies showed that the structure of PPI networks is not random. Locally, network motifs are topological patterns of nodes and edges that appear more often than expected by chance. They constitute building blocks of the global network architecture and often fulfill biological functions in cell signaling pathways such as feedback or feedforward loops (Milo et al., 2002).

The degree k of a node indicates how many edges connect the node with other nodes in the network. The degree distribution $P(k)$ gives the probability for a node to have k edges. Degree distributions of natural networks, such as PPI networks, have been

reported to follow a power-law degree distribution ($P(k) \propto k^{-\gamma}$ where γ is a constant) (Barabási and Albert, 1999; Yook et al., 2004) with many nodes forming few connections and a small number of nodes connecting a large number of other nodes. These highly connected nodes are termed hubs. Networks having this degree distribution are called scale-free. An important consequence is the small world property of PPI networks. It results in short average number of edges separating pairs of proteins.

Studies in yeast reported that hub proteins fulfill important cellular functions. Jeong et al. (2001) observed a higher number of proteins essential for growth among hub proteins. Yu et al. (2008) challenged these findings and attributed these observations to the experimental design of the initial yeast PPI screens with a higher number of essential proteins in the bait libraries (in other words, the aforementioned selection bias shapes the topology of measured PPI networks). They, in contrast, relate hub proteins to pleiotropy, a higher phenotypic diversity upon gene knock-out. Analogously, there is a debate as to whether hub proteins are more frequently among essential and disease-related proteins in humans. Proteins involved in cancer have been observed to have a higher number of PPIs (Jonsson and Bates, 2006; Wachi et al., 2005). On the other hand, Goh et al. (2007) found that the majority of disease proteins have no elevated level of interactions but essential proteins (whose orthologs in mouse lead to embryonic or postnatal lethality upon deletion) do have higher number of reported interaction partners. Again, the selection bias effect on these conclusions is difficult to estimate.

Another topological property of PPI networks is their modular organization (Barabási and Oltvai, 2004; Hartwell et al., 1999; Ravasz et al., 2002; Rives and Galitski, 2003). Modularity quantifies the strength of the network division into densely interacting groups of proteins. A measure to describe the density of a module is the clustering coefficient. It indicates for a node i the fraction of all possible edges realized among its interaction partners:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

with k_i indicating the number of interaction partners of node i and e_i the number of edges among these neighbors.

Network modules suggest that the participating proteins act closely together such as in protein complexes or cellular pathways. Accordingly, several approaches exploit the modular organization of large PPI networks to predict proteins that act together in functional subnetworks. Various network clustering tools allow the identification of groups of closely interacting proteins. Popular implementations are ClusterONE (Nepusz et al., 2012) and the Cytoscape plugin MCODE (Bader and Hogue, 2003). Usually, a large

network is scanned for modules with high clustering coefficients and more interactions formed within the module than to proteins outside the module. A maximum coefficient is achieved within a clique, which is a fully connected graph neighborhood. A graph clustering algorithm that is able to efficiently detect these structures is CFinder (Adamcsek et al., 2006).

Other studies select sets of proteins based on a certain phenotype and construct maximally parsimonious subnetworks that link these proteins (Calvano et al., 2005; Chuang et al., 2007; Said et al., 2004; Scott et al., 2005; Yosef et al., 2009). If two independent sets of proteins (e.g., membrane-bound receptors and transcription factors that respond to these receptors) are defined, edge directionality can be inferred assuming that the shortest paths between the sets represent information flow within the network (Vinayagam et al., 2011). The start nodes on these paths (i.e., the receptors) are called sources and the end nodes (i.e., the transcription factors) sinks.

A common analysis involving PPI networks is to infer the unknown function of a protein based on the known functions of its interaction partners. The underlying assumption is the guilt-by-association principle, which states that two proteins that interact likely share a common function (Oliver, 2000). This principle underlies many protein annotation tools. See for example the popular gene function prediction tool geneMANIA implemented as a web server (Warde-Farley et al., 2010) and a Cytoscape plugin (Montojo et al., 2010).

Network-level investigations have revealed intriguing topological features of disease proteins: Lim et al. (2006) showed that different ataxia disease proteins share many interaction partners. Chen et al. (2006) found Alzheimer's disease proteins more closely located to each other in the global human PPI network than expected by chance. Gandhi et al. (2006) extended this observation to other diseases. This illustrates that disease proteins are likely to interact with proteins causing the same or a similar disease. Accordingly, a variety of approaches exists that seek to predict disease proteins based on their proximity or co-clustering with known disease factors (reviewed in Barabási et al. (2011)).

Most PPI detection methods do not reveal which parts of the proteins that have been measured to interact with each other mediate the binding. The definition of protein domains and their functions (including the function of providing a binding surface for other proteins) is an important step in understanding the cellular role of proteins. Therefore, several studies aim to predict which domains mediate the binding from large sets of PPIs. These approaches usually assume that these domain pairs co-occur more often in pairs of interacting proteins than expected by chance (Deng et al., 2002; Guimarães

et al., 2006; Nye et al., 2005; Riley et al., 2005; Wang et al., 2007).

2.5 Alterations of the protein-protein interaction network in disease

As the function of a cell is a dynamic property of the cooperation between many proteins, diseases can often be attributed to the interruption of normally occurring PPIs or to the formation of novel PPIs that would not occur in a healthy cell. Consequently, research expands its focus from locating disease mutations towards a systems level understanding of how mutations can alter the function of a protein and the implications for its binding behavior. For several diseases it has been shown that changes in the PPI patterns of the disease-causing proteins contribute to disease progression (Chiti and Dobson, 2006; Giorgini and Muchowski, 2005; Ross et al., 2005; Shy et al., 2004).

An example of a broad class of diseases that is characterized by the impairment of the natural balance of the interactome are neurodegenerative diseases. These include the polyQ disorders Huntington's disease (HD) and several ataxias (Gatchel and Zoghbi, 2005). The polyQ diseases are characterized by the length expansion of a polyQ stretch over a critical length threshold. There are 86 human proteins with a polyQ stretch consisting of at least 10 glutamines in a row (allowing for one mismatch; see Chapter 5). 9 of them are known to cause neurodegenerative diseases when the polyQ stretch gets expanded over a protein-specific length threshold (Gatchel and Zoghbi, 2005). For HD the critical length is reached when the polyQ stretch in the protein huntingtin (having a wild type length of 11 to 34 residues) is expanded to over 40 glutamines.

Even though the pathomechanism of polyQ diseases remains poorly understood, the formation of insoluble protein aggregates is a key feature of all known polyQ diseases (Kopito, 2000; Ross, 1997; Tran and Miller, 1999). Biochemical and cell biological experiments have demonstrated that expanded polyQ tracts drive the spontaneous assembly of insoluble protein aggregates in disease model systems (Warrick et al., 1998), suggesting that polyQ-mediated protein misfolding and aggregation are critical for disease development. However, it remains unclear whether polyQ-mediated aggregation of proteins is the cause or the consequence of progressive neurodegeneration in polyQ diseases (Chai et al., 2002; Kuemmerle et al., 1999).

Many observed effects support the idea that the alteration of normal PPI patterns is critical for polyQ disease development:

- Interaction partners of the wild type protein are found in pathological aggregates.

For example, Chai et al. (2001) showed that aggregates formed by Ataxin-3 with expanded polyQ contain several of the wild type Ataxin-3 interaction partners.

- Many proteins found in polyQ-mediated aggregates fulfill important cellular functions. Components of the ubiquitin-proteasome system, heat shock proteins, and transcription factors have been identified (Boutell et al., 1999; Cummings et al., 1998; Mitsui et al., 2002; Nucifora et al., 2001; Suhr et al., 2001b). This supports the idea that by the recruitment of essential proteins into the aggregates important cellular processes might be blocked.

- For several interactions between a polyQ-containing and another protein an increase or decrease of the interaction strength in dependence of the polyQ length has been described, as reviewed in Li and Li (2004).

Together these observations suggest that an interplay between the loss of naturally occurring PPIs, increase of interaction propensity for wild type interaction partners as well as aberrant PPIs with the mutated protein contribute to disease progression.

3 High-confidence protein-protein interactions

3.1 Motivation

Published PPIs are collected and stored in several expert curated databases. Nevertheless, the overlap between these databases is small (Futschik et al., 2007; Lopes et al., 2011) and many reported interactions are likely false-positive observations. Few resources exist that integrate PPI data with experimental meta-data and allow users to filter out interactions that are supported only by poor experimental evidence.

The computational use of PPI datasets often requires selecting a maximum number of PPIs at a particular level of confidence. For example, the quality of a novel PPI dataset may be evaluated by its overlap with known, highly reliable interactions, whereas a statistical analysis (such as predicting domain-domain interactions from PPI data) might require a large number of interactions therefore benefiting from a less restricted set of PPIs. The flexible selection of PPI datasets at various confidence levels requires a continuous scoring scheme for PPIs reflecting the reliability of their experimental characterization.

With the objective of creating a resource containing a maximum number of interactions and allowing the selection of PPIs by experimental confidence cut-offs, we generated HIPPIE (Human Integrated Protein-Protein Interaction rEference), a scored human PPI collection integrated from multiple sources. We associated confidence values to each interaction that reflect the amount and quality of evidence supporting the interaction combining three types of information: experimental techniques used, number of studies describing the PPI, and reproducibility in model organisms. HIPPIE's scoring scheme has been optimized by human experts as well as a computer algorithm. We show that these scores correlate to the quality of the experimental characterization.

To provide a convenient tool for doing network analyses focused on likely true PPI sets by generating subnetworks around proteins of interest at a specified confidence level, we implemented a web tool together with a fully automated update routine that

regularly retrieves novel interactions from the major PPI resources and integrates them to HIPPIE.

Obviously, the selection of proteins that are selected for PPI assays is not random: Some bait and prey libraries covering a subset of the proteome are frequently used and research focuses on specific proteins, pathways or diseases. We aim to quantify the resulting bias in the integrated PPI resource HIPPIE caused by the non-uniform usage of bait proteins and describe its impact on the network topology.

3.2 Evidence-based scoring of protein-protein interactions

Expert-curated databases provide PPI data annotated with meta-information about the experiments conducted to measure the PPI. These curation efforts are increasingly standardized (see section 2.3), which allows the automated extraction and processing of relevant information. Our goal was to merge the major publicly available PPI data repositories to maximize the coverage of the human PPIs experimentally detected so far, while using the meta-information to associate each interaction with a confidence score reflecting the amount and quality of evidence supporting the interaction. While merging the different data sources we extracted information about which experimental system was used to detect each interaction and whether there were several studies in which the interaction was described. Additionally we retrieved the interaction data from PPI databases that link interactions in non-human model organisms to their human orthologs. From these different types of information (experimental systems, number of studies and reproducibility in other organisms) we calculated an overall score reflecting the reliability of each interaction.

3.2.1 Integration of human protein-protein interaction data

We retrieved and integrated all data stored in the major PPI databases listed in Table 3.1. From BioGRID we removed genetic interactions, which are generated by methods that do not require a direct physical contact of the associated proteins. We extracted binary PPI information together with annotation data required for the confidence score calculation from these resources. Additionally, we extracted the information describing whether an interaction was a direct physical interaction or whether the proteins were just identified as members of the same complex.

We identified studies that were not covered by the source databases and integrated them as well (Albers et al., 2005; Bell et al., 2009; Colland et al., 2004; Goehler et al., 2004; Kaltenbach et al., 2007; Lehner and Sanderson, 2004; Lim et al., 2006; Nakayama

Database	Size	Format	PSICQUIC	Reference
HPRD	38627	self defined	no	Keshava Prasad et al. (2009)
BioGRID	28514	PSI-MI	yes	Stark et al. (2011)
IntAct	26716	PSI-MI	yes	Aranda et al. (2010)
MINT	14739	PSI-MI	yes	Ceol et al. (2010)
BIND	1532	PSI-MI	yes	Bader et al. (2003)
DIP	1504	PSI-MI	yes	Salwinski et al. (2004)
MIPS	250	PSI-MI	no	Pagel et al. (2005)

Table 3.1: Access and curation characteristics of PPI databases integrated in the first publicly released version of HIPPIE (v1.2). The size column lists the amount of human PPIs and the PSICQUIC column indicates if the database can be programmatically accessed via PSICQUIC.

et al., 2002; Rual et al., 2005; Stelzl et al., 2005; Venkatesan et al., 2009). For these interaction sets we manually filled in the missing annotation information. For the first public version of our integrated PPI resource (HIPPIE v1.2; November 2011) we assembled 72,916 interactions from which more than 99% were associated with experimental information. Due to bi-annual updates, the number of interactions stored in HIPPIE is constantly growing. As of August 2012, HIPPIE contains 109,670 PPIs.

For confidence scoring purposes, we also retrieved data from three databases that map interactions between non-human protein pairs to their human orthologs: HomoMINT (Persico et al., 2005), I2D (Brown and Jurisica, 2005) and the PPI dataset from Lehner and Fraser (2004).

A main challenge when integrating different public PPI databases and datasets is the different use of gene or protein identifiers. We aimed at mapping all protein pairs collected in HIPPIE to Entrez Gene and UniProt identifiers. For this purpose we applied the database identifier mapping tables curated by UniProt (Apweiler et al., 2011) and the HUGO Gene Nomenclature Committee (Seal et al., 2011). We mapped all database entries to their canonical representatives and did not consider splicing forms.

3.2.2 Score calculation

We calculated a score S between 0 and 1 for each interaction reflecting the reliability of its combined experimental evidence. This score was calculated as a weighted sum of three different subscores which are s_s (a function of the number of studies in which an interaction was detected), s_t (a function of the number and quality of experimental techniques used to measure an interaction; see below for details) and s_o (a function of the number of non-human organisms in which an interaction was reproduced). Each

of these three subscores s_i were calculated with a non-linear saturating function of the form:

$$s_i(n) = \frac{2}{1 + e^{-a_i * n}} - 1$$

such that $s_i(0) = 0$ and $s_i(\infty) = 1$, where the a_i are constants that control the steepness of the function.

For subscore s_s, n is the number of different studies where the interaction was reported (number of PubMed identifiers associated), regardless of whether multiple experimental evidence was provided in each study.

For subscore s_o, n is the number of species where orthologs of the interacting proteins could be defined and were found experimentally to interact (currently *Bos taurus*, *Caenorhabditis elegans*, *Canis familiaris*, *Drosophila melanogaster*, *Gallus gallus*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, and *Sus scrofa*).

For subscore s_t, n is a sum of scores from different experimental techniques by which an interaction was verified (even if used in the same study). Most PPI databases use controlled vocabulary descriptors for these experimental techniques as defined by the PSI-MI ontology (Hermjakob et al., 2004), however for some terms we could not find an equivalent ontology term. Manual curation was used to assign a score to each PPI detection method ranging from 0 (no experiment assigned, less than 1% of PPIs) to 10. Scores and corresponding PSI-MI codes are displayed in the Appendix, Table 1. Methods that can ascertain interactions with the highest reliability, such as in vitro techniques like X-ray crystallography, were assigned the highest scores. Complementation-based assays and affinity based technologies were roughly equally scored with an average value of 5, slightly increased for those methods that are generally used in homologous, more physiological setups, such as Fluorescence Resonance Energy Transfer (FRET). Methodologies that do not directly provide evidence for interaction, such as colocalization or cosedimentation, are scored with the lowest values.

The total score S was calculated as a weighted sum of the three subscores:

$$S = w_s s_s + w_o s_o + w_t s_t$$

with $w_s + w_o + w_t = 1$.

It is important to note that our dataset includes only interactions that were experimentally verified with human proteins: no interaction received a score alone from its verification in non-human organisms. We also note that this scoring scheme does not consider computational evidence other than the definition of orthology relations from human proteins to proteins in other organisms.

3.2.3 Parameter selection

The six free parameters of the scoring formula (a_s, a_o, a_t, w_s, w_o and w_t) were optimized by performing a grid search in the parameter space. We performed the search in the range [0, 3] for the a_i and in the range [0, 1] for the w_i. We chose a step width of 0.1 for both a_i and the w_i. The step width was chosen sufficiently small such that selecting neighboring parameter combinations resulted only in small changes in the interaction scores, which decreased the probability of missing an optimal solution. Constraints were set on the weights w_i by requiring that they sum up to 1.

PPIs are sometimes reported in multiple studies. We reasoned that we could use this property to assess the performance of a parameter combination. To perform this evaluation we used the IntAct dataset (version from August 2011) consisting of 28,073 interactions (38.5% of HIPPIE). This dataset has explicit associations between studies and experiments, and the experimental information is annotated following the PSI-MI format.

The assessment of performance of a parameter set was done by successively removing each one of the 109 studies in IntAct that contain at least 10 interactions and more than 2 PPIs found in multiple studies. For each study j, we recalculated the scores of the remaining dataset, $IntAct_{red}$, found the set of PPIs described both in the study j and in $IntAct_{red}$, $\{IntAct_{red} \bigcap study_j\}$, and computed the deviation from random expectation of the number of highly scored interactions within the overlap:

$$dev_j = \frac{|scores(IntAct_{red} \bigcap study_j) > Q_3|}{\frac{|IntAct_{red} \bigcap study_j|}{0.25}}$$

where Q_3 is the upper quartile of the score distribution of $IntAct_{red}$.
To measure the overall performance of a parameter combination we chose a function f of the weighted mean of the logarithm of dev_i over all studies:

$$f = \frac{\sum_j v_j * \log_2 dev_j}{n}$$

where the weights v_j were chosen proportional to the overlap size between $IntAct_{red}$ and $study_j$ and n is the number of studies.

We found several optimal parameter combinations (several thousand optimal combinations out of more than 700,000 different parameter combinations tested) maximizing the function f (with $max(f) = 1.023$). From the equally well performing parameter combinations we chose the set of parameters that resulted in the largest spread of the distribution of scored interactions. For that purpose the scores of the entire HIPPIE

database were repeatedly calculated for each of the optimal parameter combinations and for each score distribution the interquartile range (iqr) was determined. We found that the parameter set [$a_s = 2.3$, $a_o = 1.6$, $a_t = 0.2$, $w_s = 0.6$, $w_o = 0.1$, $w_t = 0.3$] maximized both f and iqr. The optimal selection of parameters weights the reproducibility in independent studies higher than the amount and quality of experimental techniques applied and the conservation of the interaction between orthologous protein pairs.

3.2.4 Evaluation

The number of PPIs derived using different experimental system types was highly variable. HIPPIE integrates various datasets dealing with different experimental systems and thus contains a larger amount of interactions than each of those sets separately. Values for three commonly applied techniques to detect PPIs: Y2H, anti-bait coimmunoprecipitation (Coprep), and TAP are shown in Figure 3.1, which together cover 78% of the total amount of proteins in the version v1.2 of HIPPIE, but only around 50% of its interactions. Coprep and TAP share relatively many PPIs between each other (139 PPIs) compared to the other pairwise overlaps between methods. For example, TAP shares 95 interactions with Y2H despite the much higher amount of Y2H interactions as compared to Coprep. This higher overlap between Coprep and TAP in comparison with the Y2H data might reflect the similarity between the first two approaches in comparison with the latter, as Coprep and TAP are both based on antibody capture of a protein complex while Y2H is based on the reconstitution of a binary interaction inside of a heterologous system (yeast).

To illustrate the benefit of using a large dataset such as HIPPIE, we compared it with novel high-throughput PPI datasets not used for its production. We chose three high-throughput PPI datasets from the recent literature: two Y2H datasets, Y2He1 (Bandyopadhyay et al., 2010), containing 551 PPIs between 434 proteins, and Y2He2 (Wang et al., 2011), containing 3484 interactions between 2582 proteins, and a MS dataset, MSe (Behrends et al., 2010), containing 711 PPIs between 424 proteins. The coverage of the Y2He1, Y2He2 and MSe datasets by HIPPIE was of 120 (21.8%), 296 (8.5%) and 73 (10.3%) PPIs, respectively.

We evaluated the usefulness of the HIPPIE score using the three novel datasets. The HIPPIE database was divided in a high quality subset containing the top 5% highest scoring interactions (score ≥ 0.88) and a subset containing all other interactions (score < 0.88). Then, we compared the fraction of PPIs in each HIPPIE subset that was recalled by the novel dataset. If the scores are meaningful one would expect better recall of the set with high-confidence scores.

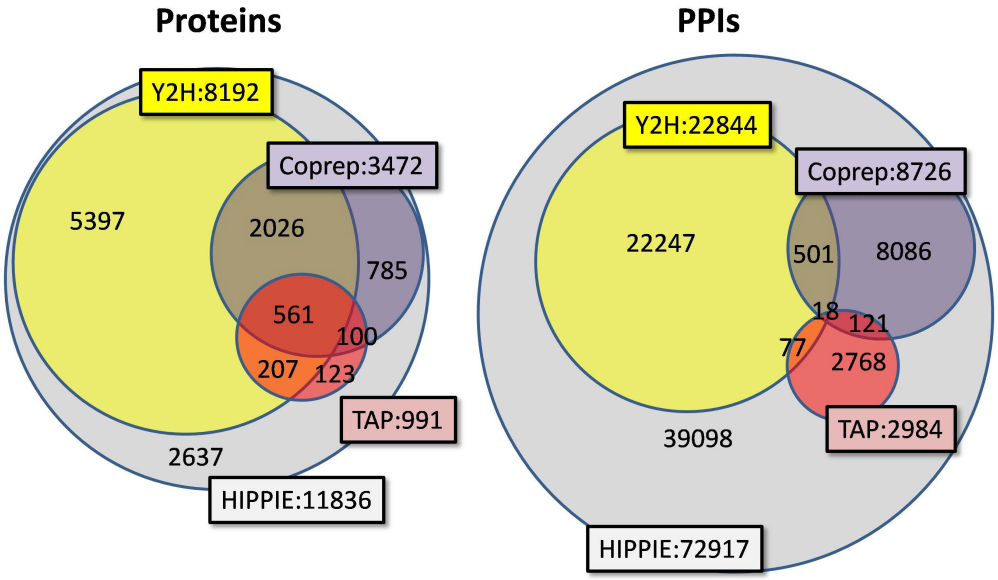

Figure 3.1: Coverage of HIPPIE v1.2 and overlap by three technique-specific datasets.

To measure the recall of HIPPIE by an external dataset of PPIs we considered that some PPIs from HIPPIE may not be detectable by the experimental setup used to produce the external dataset: in case of the MSe, the set of bait proteins was given, so all interactions from HIPPIE in which at least one of these bait proteins was participating were considered to be detectable PPIs. For the Y2H experiments, we chose as detectable PPIs all interactions in HIPPIE where both interacting proteins were also found to participate in interactions of the experimental set. The number of detectable PPIs and the recall (Table 3.2) were used to calculate chi-squared tests to assess the significance of the differences in recall between high and low confidence HIPPIE subsets. The high quality subset had the largest overlaps in percentage with the PPIs of the novel datasets and these overlaps were significant (p-values of 8.2e-12, 2.2e-16 and 9.9e-14 for Y2He1, Y2He2 and MSe, respectively) suggesting that the PPI score correlates with experimental reproducibility.

To compare the performance of the HIPPIE score with the confidence score of MINT, we contrasted the recall of detectable PPIs for both scoring schemes (Table 3.2). For all tested studies, the recall of the detectable interactions was larger for PPIs with high HIPPIE scores than for PPIs with high MINT scores. Accordingly, the fraction of PPIs found in the studies among the low scoring detectable PPIs was higher when the MINT scoring scheme was applied than for the HIPPIE score.

		HIPPIE		MINT	
		HC	**Other**	**HC**	**Other**
Number interactions		3818	69319	945	17555
Y2He1	**Detectable PPIs**	142	944	40	296
	Overlap (recall)	40 (0.28)	80 (0.08)	9 (0.23)	30 (0.10)
Y2He2	**Detectable PPIs**	821	9135	183	2884
	Overlap (recall)	98 (0.12)	198 (0.02)	17 (0.09)	90 (0.03)
MSe	**Detectable PPIs**	71	1060	13	322
	Overlap (recall)	20 (0.28)	53 (0.05)	3 (0.23)	17 (0.05)

Table 3.2: Coverage of HIPPIE v1.2 and MINT (version from 06/02/12) by novel datasets. Both databases are split into a high confidence set (HC) containing the 5% highest scoring interactions and a set consisting of all other PPIs. Only interactions that could potentially be detected by the design of the study are considered and the recall of these interactions is calculated. For all studies tested, the recall of the high-confidence detectable interactions was larger for HIPPIE than for MINT.

3.3 Impact of study design on network topology

It is a common network analysis task to identify hub proteins that are characterized by many interaction partners (see Chapter 2 for several examples) or to draw general conclusions from the degree distribution of sets of proteins (in Chapter 5 we will relate degree characteristics of a group of proteins to their function). These analyses are usually done on integrated PPI networks such as HIPPIE. We were wondering to which extent the fact that some proteins are chosen as baits more often than others will bias the observed degree distributions. An important question would be if proteins with many interaction partners are necessarily true hub proteins or if they are just proteins that are studied more intensively.

Bait proteins are labeled as such in the manually curated PPI database IntAct, which was used among others to assemble HIPPIE. We retrieved this information for 47,909 pairwise interactions and assembled a list of 4233 bait proteins associated to the number of studies in which they were examined for interaction partners. This list does not cover all PPI experiments that contribute to HIPPIE so it only gives a lower bound for rates at which proteins have been studied. However, due to the large number of annotated experiments, we expect that this analysis gives a good approximation of the frequencies with which proteins are screened for interaction partners.

We found that the bait usage distribution can be significantly well fitted to a power

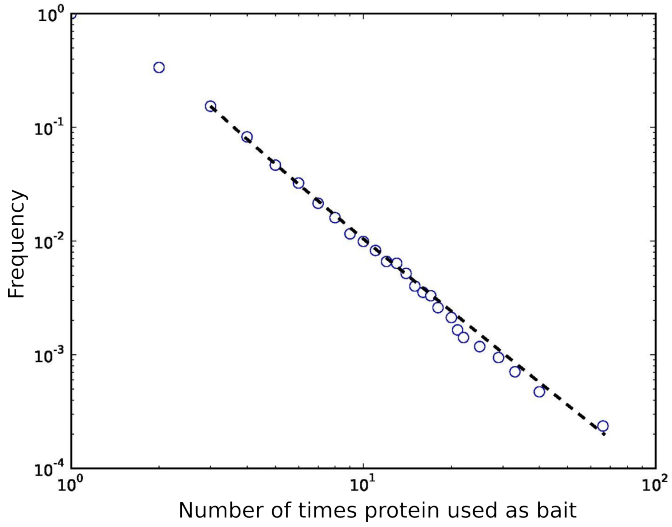

Figure 3.2: Bait usage statistics. The number of times a protein has been screened for interaction partners can be approximated with a power-law distribution.

law distribution with exponent $\gamma = 3.04$ for number of baits ≥ 3 (Figure 3.2) using the method of Clauset et al. (2009) (for estimating parameters and associated uncertainty). As it is characteristic for a power-law distribution, we observed several proteins being used many times (for example, TP53 was used as a bait in 66 PPI screens) while the majority of proteins have been used as a bait in only one study. Using ConsensusPathDB (Kamburov et al., 2011) we tested the bait proteins for enrichment of functions and pathways considering only categories enriched with a q-value (the false discovery rate adjusted equivalent to the p-value) below 0.01. In accordance with a previous study (Futschik et al., 2007) that investigated functional categories enriched among entire networks, we found a strong enrichment of proteins with nuclear localization, involved in cell cycle and metabolism ($q < 0.0001$) among the proteins used as baits. When calculating the enrichment of functional terms and pathways among 197 proteins frequently used as a bait (more than four times) relative to that of the full bait list, most strongly enriched were "pathways related to cancer" ($q < 10^{-32}$). While the enrichment of nuclear proteins in the entire bait set might be caused by a technical detection bias of the still predominantly used Y2H assay, which requires nuclear localization of the bait and prey proteins, the strong enrichment for cancer pathways in the frequently studied bait set clearly indicates a selection bias towards proteins with high biomedical relevance.

To test if the intuition holds that intensively studied proteins have more interaction partners than less intensively studied proteins, we calculated the correlation between the degree and the number of experiments in which the protein was used as a bait.

Indeed, we observed a positive (Pearson) correlation ($r = 0.552$) that is not very high but significantly deviates from random expectation ($p < 10^{-16}$). Also, this correlation is higher than the previously described correlation between protein abundance and degree, ranging for different TAP/MS networks from 0.21 to 0.46 (Ivanic et al., 2009). This indicates that it might be problematic to consider proteins with many interaction partners in integrated PPI networks as hubs without controlling for the non-uniform bait usage distribution. In Chapter 5 we will address this problem and present two strategies to address the impact of selection bias on the degree distribution.

Next, we asked if pairs of intensively studied proteins are more likely reported to interact. For this purpose, we compared all protein pairs where both proteins were used in at least five studies and observed that a total of 1770 interactions was realised among the 19503 possible combinations (including self-interactions), which gives a rate of 9.1%. In comparison, among all possible combinations of proteins used only once as a bait a much lower frequency of 0.2% were observed to interact. To test if this observation was only caused by the higher degree of well studied proteins (assuming that two proteins with more interactions have a higher chance to interact with each other under random conditions), we randomly sampled protein sets of the same size as the set of the most highly studied bait proteins while preserving the degree distribution of the proteins in the set and counted realized interactions between proteins within these sets. Doing this, we found the interaction number within the set of highly studied bait proteins significantly larger than expected by chance ($p < 0.001$) and therefore conclude that this cannot be solely attributed to the higher degree of well studied proteins.

One explanation for the high association of well studied proteins might be the functional bias within these protein sets. Since many proteins share pathway membership in cancer related pathways, they are more likely to interact. On the other hand, this high number of internal associations can be caused by the biased selection strategy in combination with experimental methods that are associated with high error rates: if certain protein pairs are repeatedly tested, a noisy assay with high false-positive rates will report their interaction sooner or later.

We examined the score distribution of interactions within the set of highly studied bait proteins and among proteins only studied once (Figure 3.3). We found the score distribution of the interactions among highly studied proteins significantly larger ($p < 10^{-16}$) and, interestingly, having a larger variance ($s^2 = 0.019$) as compared to the proteins not frequently studied ($s^2 = 0.011$). Also, the former distribution has a rather bimodal shape. Closer inspection of interactions with confidence scores in the lowest and highest quartile revealed that more than 85% of the interactions with a high score

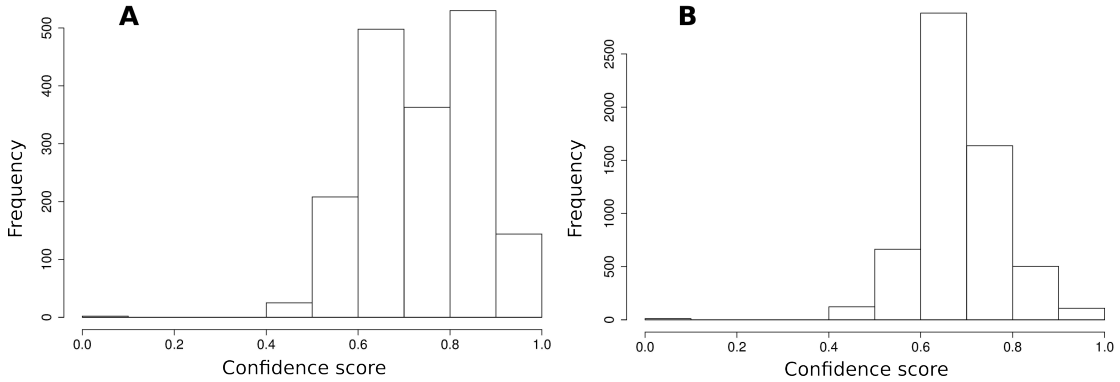

Figure 3.3: Confidence score distributions for interactions between intensively (A) and rarely (B) studied proteins.

among proteins frequently screened were detected with multiple low throughput methods (such as affinity chromatography technique or reconstituted complex) while the low scoring interactions were usually poorly annotated or had experimental descriptions that were assigned low weights by our experimental assignment scheme. Due to the design of the confidence score, the high fraction of reliable methods among the high scoring interactions is not surprising but, again, illustrates the necessity for evidence-based filtering that might be able to reduce the effect of repeated testing of the same proteins with erroneous methods.

In summary, the biased selection of proteins for interaction screening has a significant impact on the network's degree distribution, hampering strategies that identify hub proteins using only the reported interaction amount. Additionally, pairs of intensively studied proteins are more often reported to interact than expected by chance.

3.4 Implementation of the HIPPIE web tool

3.4.1 Design

We implemented a web interface that allows one to query and analyze the PPI data stored in HIPPIE. The web layer of HIPPIE was implemented in PHP. For graph visualization it embeds the network viewer Cytoscape Web (Lopes et al., 2010). The PPIs and associated meta data are stored in a MySQL database. Network analyses on the PPI data (as will be described in Chapter 4) are performed using Python and the graph algorithm library NetworkX (Hagberg et al., 2008). Several components of the query interface (such as the tree structure for the selection of functional filters described in Chapter 4) are implemented in JavaScript making use of the libraries jQuery and dynatree. The HIPPIE

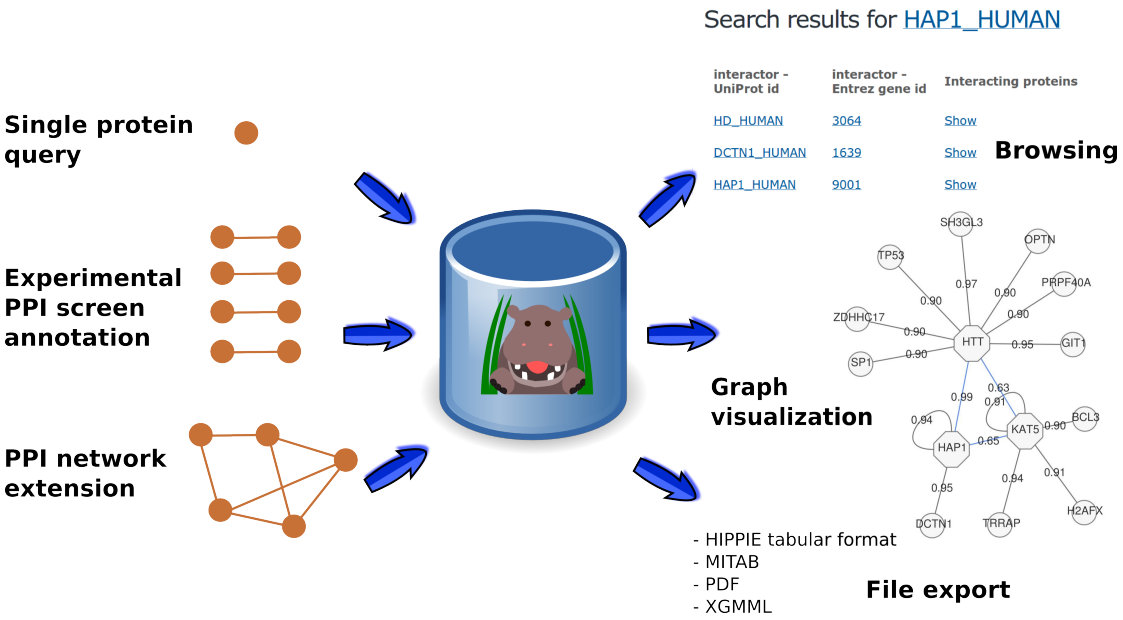

Figure 3.4: Summary of HIPPIE query options and the various ways to output the generated networks.

update routine is implemented in Java. Via PSICQUIC (Aranda et al., 2011) it accesses those source databases that implement a programmatic interface (see Table 3.1). It automatically integrates the retrieved PPI data, rescores the entire PPI repository based on the updated evidence records and releases bulk download files of the newly generated HIPPIE version. Additionally, we implemented several tools in Java that allow one to execute some of HIPPIE's analysis tasks locally for larger input sets (see following section). The web tool can be found at http://cbdm.mdc-berlin.de/tools/hippie.

3.4.2 Query options

The HIPPIE web tool allows one to access and query the PPI data in different ways. The query and output options are summarized in Figure 3.4. In the most simple case, HIPPIE can by queried using a single gene symbol, Entrez gene ID or UniProt identifier (ID and accession) (see the protein query interface in Figure 3.5). On the result page, a confidence score is listed with each interaction partner of the query protein and detailed information about the evidence contributing to the confidence score can be accessed. Links to the original studies are provided. For each interaction a link is given that generates a new HIPPIE query with the interaction partner as an input.

A typical problem after the generation of experimental results that produce a list of

genes, proteins and/or interactions between them, is the evaluation of the results in relation to the already known PPI data. For example, a researcher may have obtained proteomics data for a few proteins of interest and wants to evaluate the novelty of the interactions, or the possible relation of the interactors with a disease-related protein of interest. Two query options facilitate this analysis. (a) A list of interactions can be uploaded to HIPPIE and for each interaction it is indicated if the PPI has been reported in the literature before (i.e., if it is found in HIPPIE) and for known interactions the confidence score is reported. (b) Additionally, HIPPIE can be queried with a set of proteins and/or interactions between them from which a network of known data around the proteins of interest is constructed. The online tool will identify interactions between the proteins submitted (layer 0 network), or their interactors not contained in the query set (layer 1 network). The computation of networks with more layers might be lengthy if hundreds of protein partners have to be analysed. For this we provide a Java command line tool (available from http://cbdm.mdc-berlin.de/tools/hippie and also deposited at the SourceForge open software archive: https://sourceforge.net/projects/hippiecbdm) that will perform the computation on the user's local machine for large input sets or neighbours of neighbours. A confidence threshold to control the reliability and size of the constructed network can be also applied. Additionally, we provide a filter option for the PSI-MI interaction type annotation provided by most of the source databases of HIPPIE. This feature allows for selecting direct physical interactions from HIPPIE. The resulting HIPPIE subnetworks can then be displayed in tabular format, exported from HIPPIE to a text file for further analyses or can be visualized using the tool Cytoscape Web (Lopes et al., 2010), which has been integrated into HIPPIE. In the visualization mode several types of information associated with proteins and interactions are visually encoded. For example, the color of edges in the generated network indicates which interactions have been uploaded and if they are present in HIPPIE (in Chapter 4 we will present an example of HIPPIE's visualization options).

The web site also offers the entire HIPPIE dataset for download in two different formats: in PSI-MI TAB 2.5 format as defined by the Protein Standard Initiative (Hermjakob et al., 2004) and in our own tab delimited flat file format. Detailed usage instructions and description of the features of the HIPPIE web tool are available at the HIPPIE web page (http://cbdm.mdc-berlin.de/tools/hippie).

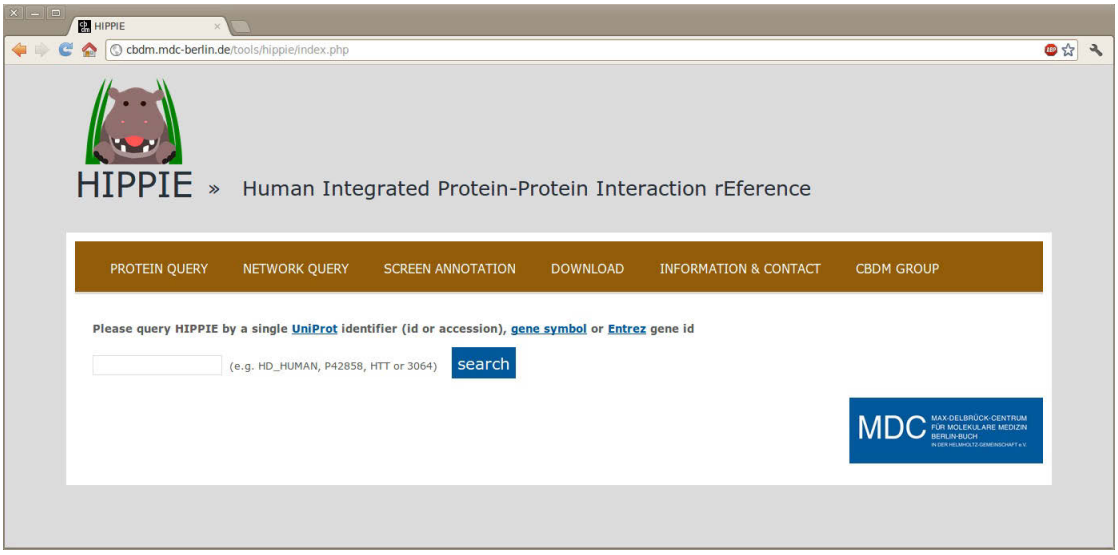

Figure 3.5: Protein query page of the HIPPIE web tool.

3.5 Discussion

In this chapter, we presented a method to score PPIs based on the experimental evidence supporting the interaction as well as its implementation in HIPPIE, an integrated dataset of human protein interaction data with associated confidence scores. This resource has been created for researchers that need to use the complete knowledge of human protein interactions. This is required in systems biology studies and in the evaluation of high-throughput results (e.g., novel PPI datasets) that require contrasting results with interactions selected for a particular level of reliability.

HIPPIE currently (version 1.4; August 2012) integrates 109,670 interactions from several public PPI resources scored according to confidence. For comparison, the complete human interactome map has been estimated to contain between 130,000 and 260,000 interactions (Hart et al., 2006; Venkatesan et al., 2009). Considering the expected high frequency of false-positives among the low scoring interactions, this suggests that our knowledge of the human interactome is still incomplete. Nevertheless, producing a large collection of integrated PPI data is critical for its usability because novel high-throughput PPI datasets often contain just a few hundred PPIs and might have little overlap with smaller existing PPI resources integrated in HIPPIE.

HIPPIE has been used for the evaluation of existing novel PPI datasets showing that it increases their coverage over individual resources and that its scoring scheme correlates with the ability to find a PPI in experimental data not included in the database

(Table 3.2).

Several resources have been created that, like HIPPIE, integrate PPI data from multiple sources but do not have a focus on distributing a scored dataset, while offering excellent tools to examine the evidence supporting each PPI. Examples include iRefWeb (Turner et al., 2010) and UniHI (Chaurasia et al., 2007). STRING (Szklarczyk et al., 2011) offers a confidence score weighting functional associations but does not focus on experimentally verified interactions. MINT (Ceol et al., 2010) provides an evidence-based confidence score similar to ours. However, as a manually curated resource it covers only a fraction of the interactions stored in HIPPIE. We also showed that our score optimization leads to a better correlation with experimental reproducibility than for the MINT score.

We are aware that any assignment of reliability scores to experimental techniques necessarily reflects the individual beliefs of researchers. We tried however to base our selection of parameters and weights in the scoring formula on objective criteria by optimizing the performance of our scoring scheme to assign high values to reproducible interactions. For researchers who nevertheless wish to modify either the selected parameters or the scores assigned to the different techniques we offer a tool at the HIPPIE homepage that allows the rescoring of interactions in HIPPIE using a different set of parameters.

Investigating the effects of integrating a large number of studies (currently 27,788) on the usage statistics of proteins, we show that the bait usage distribution does not converge towards a uniform distribution. On the contrary, a few proteins have been screened multiple times and the majority only once or never. This bias has both functional and topological implications. The here described correlation with the degree distribution is stronger than the previously reported correlation between protein abundance and degree (Björklund et al., 2008; Ivanic et al., 2009). Few studies exist that aim at controlling for the selection bias (an exception is Dickerson et al. (2010)). Our analyses suggest that previous studies that draw conclusions from the degree distribution of integrated PPI networks (for example those of Wachi et al. (2005) and Jonsson and Bates (2006), which described a higher number interaction partners for cancer proteins based on an analysis of integrated networks) need to be carefully re-examined. In Chapter 5 we will use the HIPPIE network to study the properties of polyQ-containing proteins, some of which are involved in neurodegenerative diseases and therefore have been extensively studied. In this analysis, we will demonstrate how important it is to take the here described selection bias into account.

Providing a tool to select interactions based on their amount of experimental evidence

helps to provide PPI networks of higher reliability. Still, the problem remains that even interactions that are reproducible under artificial experimental setups might never or only under certain conditions be realised. In a recent study (Lopes et al., 2011) we showed that applying tissue expression filters to subnetworks of HIPPIE around human proteins that directly interact with viral proteins, we strengthen the enrichment of pathways known to be involved in the respective disease caused by the virus. In Chapter 4 we will extend this idea and show how the incorporation of functional and expression information into PPI networks leads to the detection of PPIs that are not only experimentally more reliable but also show a high relevance to human disease.

3.6 Contributions

This chapter is a modified and extended version of Schaefer et al. (2012a). The described integration of PPI data, design and optimization of a scoring formula and implementation of a web tool were done by me. The evaluation of PPI detection methods (Appendix A, Table 1) was done by our experimentally working collaboration partners (Pablo Porras and Erich Wanker). The original paper was written by Miguel Andrade and me. The study of the network topology and the performance comparison with existing approaches (which are not part of the publication) were performed and written by me.

4 Context-specific protein-protein interactions

4.1 Motivation

The advent of high-throughput techniques to measure and perturb molecular species in a systematic way has enabled researchers to assess the different layers of cellular metabolism under different experimental conditions. In Chapter 3 we integrated a large PPI network and developed a strategy to deal with the high error rates associated with PPI data. Another major drawback of these data is that the artificial expression systems used to reconstruct PPI networks do not take into account two of the many factors that are essential to understand the biology of the cell: first, the time-point at which the proteins are expressed (e.g., cell-cycle or developmental stage) and second, the tissue or intracellular compartment where the proteins are expressed or located (different organs and tissues have very specific protein compositions). Therefore, two proteins may be reported as interaction partners, although they are expressed in different tissues or at different time-points. While high-throughput studies acknowledge these caveats, PPI databases collect these data without mechanisms explicitly directed to discern the biological plausibility of a reported interaction. Therefore, the selection of proteins expressed in a specific cell type or compartment would allow the generation of subnetworks that would more realistically represent biological processes in the respective cell types or cellular compartment.

Several attempts have been made to investigate the tissue-specific binding behavior of single proteins and the spatio-temporal dynamics of PPI networks (Agarwal et al., 2010; Bossi and Lehner, 2009; de Lichtenberg et al., 2005; Han et al., 2004; Taylor et al., 2009; Wen-hsien et al., 2009). In a recent study evaluating the characteristics of publicly available PPI databases, we demonstrated that the use of subnetworks (which include only interactions of proteins expressed in the same tissue) identifies potential mechanisms or pathways that would remain obscured if the complete PPI database was used (Lopes et al., 2011).

In addition, many proteins have multiple functions, carried out in cooperation with distinct sets of interacting partners. Networks of interacting proteins with coherent function have been termed context networks (Rachlin et al., 2006). Here, we adopt this notion of context and extend it to PPIs or networks of proteins being expressed in the same tissue or cooperatively transmitting signal flow. There is a lack of studies testing systematically the potential of adding context information to PPI networks in recovering meaningful PPI subsets and, although there are a few approaches that allow to add expression or functional information to PPI data (Chowdhary et al., 2012; Lee et al., 2009; Yang et al., 2008), convenient methods for the creation of such context-specific subnetworks are generally missing. In this chapter, we introduce an approach to add context to PPI networks using annotations and relations between the interacting partners and demonstrate that context-specific PPI networks are enriched in high-confidence interactions. We use this approach to investigate how the proteins of the human influenza virus interfere with the immune response of the host cell in a tissue-specific manner, finding novel potential regulators of influenza virus pathogenicity, and to study the brain-specific signaling pathways that play a role in Alzheimer's disease, identifying a pathway involving the altered phosphorylation of the Tau protein. Thereby, we illustrate how the addition of context to PPI networks can guide researchers in the discovery of meaningful interactions and pathways, which would otherwise be obscured by the vast amount of irrelevant (for a specific question) and partly erroneous amount of PPI data.

4.2 Context-specific and directed protein-protein interaction networks

We inferred context information for all interactions in the human PPI database HIPPIE described in Chapter 3. In a first step, we associated all proteins in HIPPIE with the following attributes: tissue-expression, GO biological process and cellular compartment, and inferred annotations for the MeSH categories disease and tissue. Proteins were associated with tissues (based on their gene expression profiles retrieved from BioGPS (Wu et al., 2009) and using the method defined by Lopes et al. (2011)) or annotated as housekeeping (using a list from Eisenberg and Levanon (2003)). Next, associations with biological processes and subcellular locations were determined according to the EBI GO annotation (release from October 28, 2011; reduced to GO slim terms) (Dimmer et al., 2012), and to MeSH terms belonging to "Diseases" (class C) or "Tissues" (class A10) that annotate the biomedical references associated to them in MEDLINE (release 2012; gene2pubmed at NCBI ftp site).

We then inferred context associations to the PPIs according to the annotations of the interacting proteins and taking into account the hierarchical structure of GO and MeSH terms. We associated an interaction with a tissue when both interactors are expressed in the same tissue (e.g., "lung"). Given a term of a functional ontology, we associated an interaction with this function when both interactors are annotated with either the given functional term or with children of it in the hierarchy of the ontology. For example, the GO term "transport" would be associated with an interaction between a protein annotated as involved in "vacuolar transport" and another protein annotated as involved in "nucleocytoplasmic transport". We excluded the rather unspecific top-level terms "biological process", "cellular component" and "cell". Additionally, we ignored categories that are associated to less than 20 interactions.

We implemented filters for the HIPPIE web tool that allow to generate context-specific subnetworks by selecting the respective category (see Figure 4.1). Additionally, HIPPIE implements the prediction of information flow and edge directionality. This is done assuming that signal pathways follow the transmission of information through interacting proteins starting in cell surface receptors that collect external cues and ending in transcription factors as final effectors on gene regulation, following Vinayagam et al. (2011). All pairwise shortest paths between proteins annotated with the GO term "receptor" and "sequence-specific DNA binding transcription factor activity", respectively, in the UniprotKB (Magrane and UniProt Consortium, 2011) were computed. An edge of the network was considered to be directed if at least one shortest path goes through that edge. The direction of the path (from source to sink) determined the direction of the edge. Edges with conflicting orientations of passing paths were not assigned directionality.

The HIPPIE web tool allows users to specify generic source and sink sets (instead of receptors and transcription factors) between which the shortest paths are computed in output networks that result from a HIPPIE query. We do not consider edge weights and, hence we are able to determine each shortest path in linear time via a breadth-first search.

Overall, we were able to associate context to 95% of the more than 100,000 interactions of the current version of HIPPIE (only considering function and tissue expression, not the edge directions). Interactions for which we inferred or collected annotations had significantly better experimental evidence (Figure 4.2). This suggests that annotated interactions might have higher biological significance than non-annotated ones.

We observed that more specific context categories were associated to interactions with higher experimental reliability: while the confidence scores of interactions with

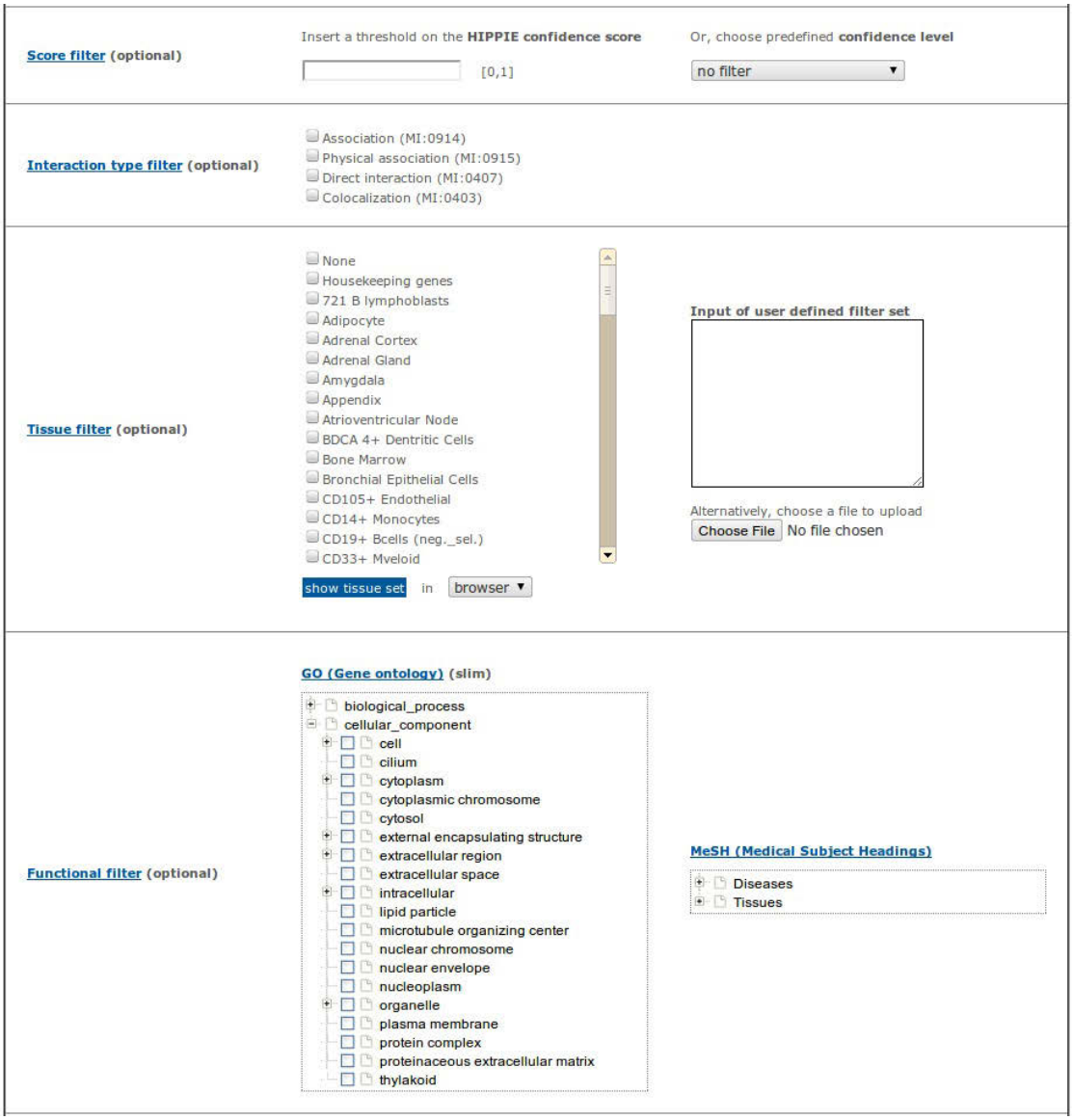

Figure 4.1: Generation of context-specific PPI networks with the HIPPIE web tool. Various query options allow to filter for edge annotations.

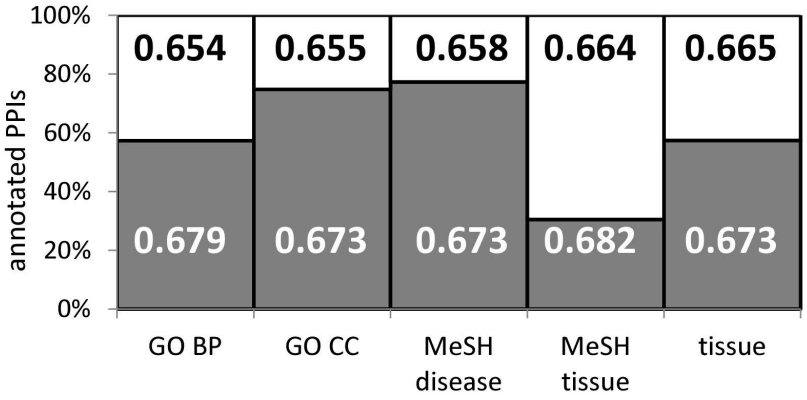

Figure 4.2: Context-associated interactions have in average a higher confidence score than non-annotated interactions. The numbers in the bars indicate the mean experimental score of the non-annotated fraction (above, black font) and of the annotated fraction (below, white font), respectively. All mean-score differences between annotated and not annotated interactions were significant ($p < 0.001$; Mann-Whitney-test).

rather unspecific and ubiquitous terms resemble the overall confidence score distribution, interactions with highly specific terms usually have a higher than average confidence score (Figure 4.3). For example, the 43,372 interactions associated with the GO category "cytoplasm" (of depth 1 in the GO hierarchy) have an average confidence score of 0.675 as compared to the average of 0.670 over all interactions. On the other hand, the 159 interactions associated with the (depth 3) GO category "ribonucleoprotein complex assembly" have a high average confidence score of 0.754. We observed a similar tendency for more specific MeSH terms to have a higher experimental reliability.

To demonstrate that our automated context association approach allows identification of relevant interactions, we tested if networks of interactions of our inferred MeSH-based disease-annotation are enriched in well-known disease proteins. Therefore, we repeatedly generated disease-context networks around a set of canonical disease proteins and examined if these networks included other known disease-related proteins. As a canonical disease protein specification, we retrieved the manually curated UniProt Knowledgebase disease protein annotation. For each of the canonical disease proteins, we generated two types of networks: (a) disease networks consisting only of interactions of the disease proteins that we had associated with the equivalent MeSH disease term and (b) unfiltered PPI network consisting of all interactions of the disease protein from HIPPIE. We did this for all disease proteins where the disease was associated with at least two disease

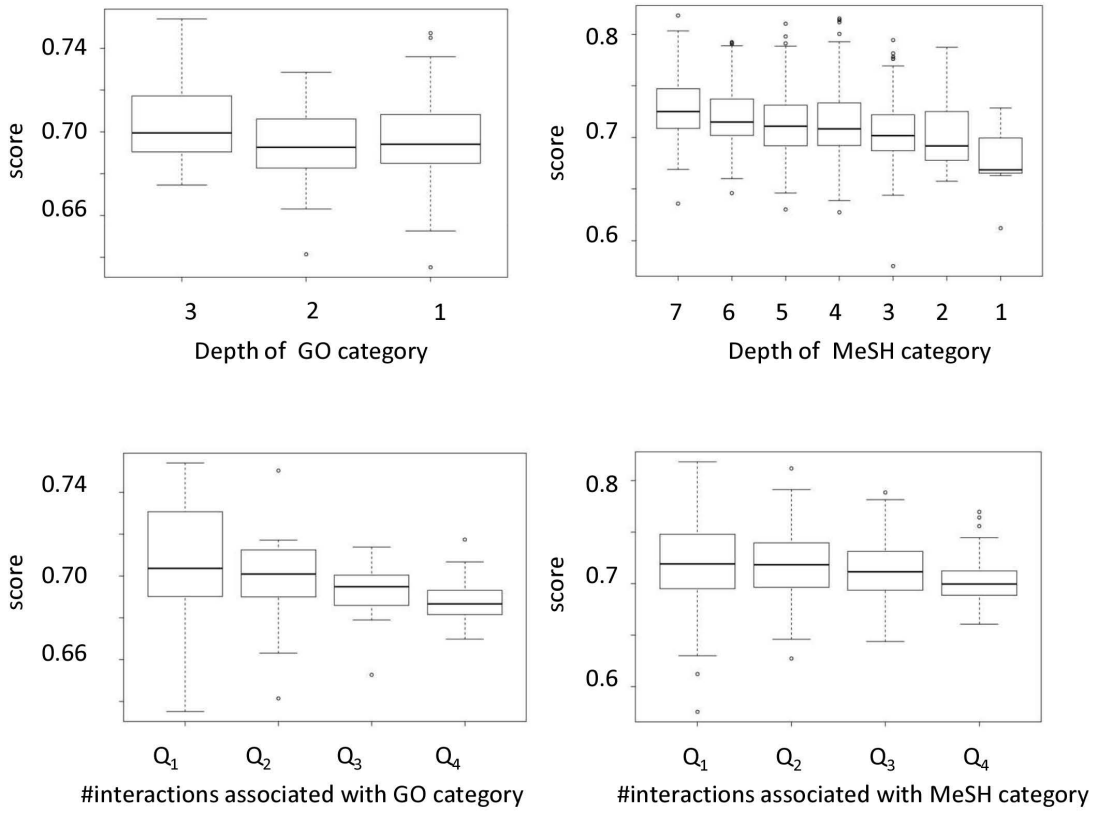

Figure 4.3: More specific edge annotations are associated with higher experimental confidence scores. The box plots visualize the distribution of experimental scores of PPIs associated with GO (left) and MeSH (right) term categories. (Top) The scores for GO and MeSH terms decreased generally for less specific terms (the only exception was GO terms depth 2, which was associated with interactions of a lower mean confidence as compared to GO terms depth 1). (Bottom) GO and MeSH terms were subdivided in quartiles according to the number of interactions annotated for each category (from low, Q1, to high number, Q4). The scores decreased for terms associated to higher numbers of interactions.

proteins in UniProt and at least two interactions that we had associated with this disease. To quantify the enrichment of disease proteins in these networks, we repeatedly calculated the F1 score, the harmonic mean of precision and recall ($F1 = \frac{2*precision*recall}{precision+recall}$). A one-sided Mann-Whitney-test comparing the distribution of F1 scores between the disease networks and the non-filtered networks indicated that the F1 scores for the disease networks were significantly larger ($p < 0.05$) proving an enrichment of disease proteins in the disease filtered networks (without losing sensitivity by removing disease proteins in the filtering step). The mean precision on the filtered networks was 0.47 and on the unfiltered networks 0.21. The mean recall for the filtered networks was 0.14 and for the unfiltered networks 0.15. This illustrates that in return for a small decrease in recall the precision can be more than doubled by applying the MeSH disease filter.

We then investigated the potential of edge directionality inference based on the shortest paths between membrane-bound receptors and transcription factors through the PPI network to recover known pathways. We retrieved pathway annotations (extracted from WikiPathways download March 29, 2012) and computed the shortest path through HIPPIE between all pairs of receptors and transcription factors within the same pathway (excluding only pairs that directly interact or could not be connected by any path). We counted the number of proteins of each pathway found on the shortest path (excluding the source and the sink node between which the shortest path was computed). We found for 3163 of the 5063 pairs that this approach correctly identified proteins of the selected pathway. The mean precision (the fraction of proteins on the paths that indeed belonged to the correct pathway) over all combinations of receptors with transcription factors was 0.20. The mean recall (the fraction of the pathway that was recovered by considering the paths between one receptor and one transcription factor) was 0.02.

To assess if the agreement between shortest paths and canonical pathways was larger than expected by chance, we generated a background distribution by computing repeatedly the shortest paths between a receptor and a transcription factor from different pathways and computed the overlap between the proteins on the shortest path to either the transcription factor- or the receptor-containing pathway. We found that the overlap distribution was significantly higher when the receptor and the transcription factor were members of the same pathway ($p < 0.001$; Mann-Whitney-test) proving the potential of shortest paths to recover the signal flow between transcription factors and receptors.

We wondered if we could further increase the overlap between the shortest paths and the canonical pathways by filtering the networks for tissue expression. To associate pathways with tissues, we determined for each pathway which tissues were more than two-fold enriched among the genes of the pathway (using again the earlier described

association between tissues and genes). Inspection of the tissues enriched among proteins forming a pathway revealed that in many cases they indeed reflect plausible locations for pathway activity. For example, immune response pathways were enriched among blood cells, and pathways associated with neurodegenerative diseases and addiction were enriched in brain-related tissues.

We repeated the computation of shortest paths linking receptors to transcription factors in tissue-specific networks for all combinations of pathways and tissues and for all pairs of receptors and transcription factors that were expressed in the respective tissue. Indeed, we observed an increase of the mean precision to 0.24 (as compared to a precision of 0.20 for the unfiltered networks), which shows that we could enhance the agreement between shortest paths and canonical pathways by computing the shortest paths in tissue-specific networks. The recall increased from 0.02 to 0.03, which is still a low value but not surprising since many pathway-related proteins were not present in the considered tissue-specific networks and, hence, could not be detected. Again, the amount of pathway proteins on the tissue-specific shortest path between receptors and transcription factors from the same pathway was significantly larger as compared to shortest paths between receptors and transcription factors from different pathways ($p < 0.05$).

To further investigate if the described context-associations can help to extract pathway information from networks, we compared the frequency of protein pairs being members of the same pathway (as defined by WikiPathways) among tissue-specific PPIs (both proteins where required to be co-expressed in at least one tissue) and to the frequency among PPIs between proteins that are not expressed in the same tissue. We observed that interacting protein pairs that are expressed in the same tissue are indeed more likely to be in the same pathway as compared to interacting protein pairs that are expressed in disjoint sets of tissues ($p < 0.001$). This, again, demonstrates that the annotations have captured properties related to pathways and suggests that the filtering helps revealing pathway information.

In the next sections we use the context-associated PPI network to obtain novel insights into the mechanisms of human disease: we perform a targeted study of the PPI network surrounding the human proteins that interact with influenza virus proteins to find potential regulators of viral pathogenicity, and we explore the question of whether and how altered protein phosphorylation might be a cause of Alzheimer's disease.

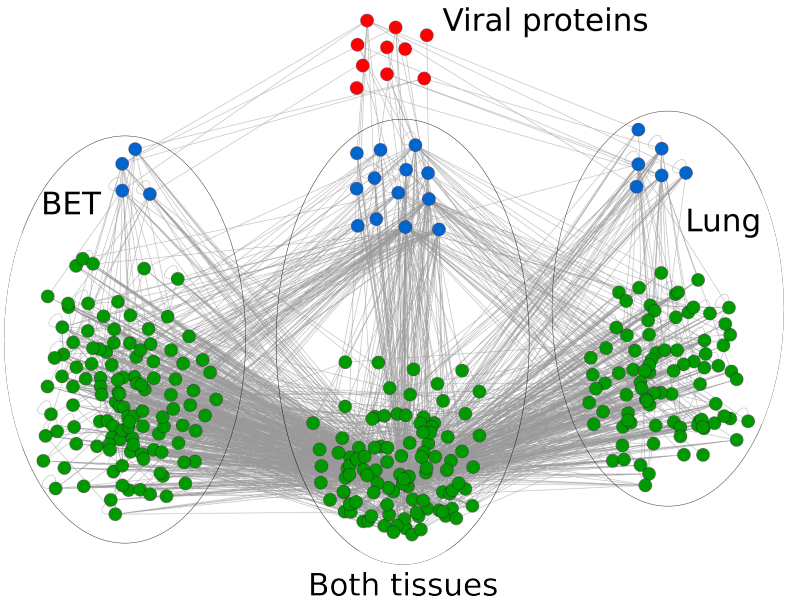

Figure 4.4: Tissue-specific PPI subnetwork of human proteins interacting with influenza virus proteins. (A) Influenza proteins (red) interact with 23 first layer host proteins (blue). These first layer proteins have interaction partners that are specific for the bronchial epithelial tissue (BET) subnetwork, for the lung subnetwork or are shared between both subnetworks (all in green).

4.3 Context-specific influenza host factor networks

We analyzed PPI data of human proteins that interact with influenza virus proteins. Influenza viruses infect bronchial epithelial tissue and many cell types in the lung, sometimes resulting in viral pneumonia (Fields et al., 2007). We started by obtaining a list of 87 human proteins that have been shown to interact with at least one influenza virus protein in a previous study (Shapira et al., 2009). From this list, we observed that 23 proteins were expressed in bronchial epithelial tissue (BET), in whole lung, or in both tissues - we refer to these proteins as first layer host factors. We created the second layer by filtering tissue-specific proteins (expressed in BET or whole lung) that interact with members of the first layer (Figure 4.4). Together, the first and second layers compose the tissue-specific PPI subnetworks.

Next, we analyzed the BET- and lung-specific PPI subnetworks using pathway enrichment analyses, and found both similarities and differences in the cellular functions of each. We performed pathway enrichment analysis with ConsensusPathDB (run on August 30, 2012; (Kamburov et al., 2011)). We used a cut-off of 0.05 on the q-value (the false discovery rate adjusted equivalent to the p-value). The background control for the

tests was the complete list of proteins annotated as expressed in the given tissues (and with PPI information in HIPPIE).

Both subnetworks showed enrichment for processes related to programmed cell death and eukaryotic translation. These results are consistent with functions known to be activated or disrupted by influenza virus infection (Ehrhardt et al., 2010; He et al., 2010; Ludwig et al., 2006). In addition, proteins in the BET subnetwork exhibited a stronger signature in processes involved with transcriptional regulation, sumoylation, and the regulation of mRNA stability (in particular, the stability of AU-rich element-containing mRNAs). Although these processes tend to be associated with general housekeeping functions, we point out that many cytokine and interferon mRNAs contain AU-rich elements (Khabar, 2005). This observation suggests, hypothetically, that influenza virus proteins may function to dysregulate cytokine mRNA stability in BET, a function that could impact influenza virus pathogenesis through modulation of immune cell infiltration and function. In relation to sumoylation, it has been noted recently that influenza virus can gain protein functionality during infection by interacting with the sumoylation system of the host cell (Pal et al., 2011). On the other hand, the lung subnetwork was uniquely enriched for processes related to cell-substrate adhesion (pathway "signaling events mediated by focal adhesion kinase"). Because cell adhesion is important for maintaining cellular viability and epithelial barrier function, it is possible that influenza virus protein-mediated interference with this process could impact both the amount of virus-inflicted damage upon the lung and the dissemination of influenza virus into extra-pulmonary sites.

Cells respond to influenza infection by producing cytokines and chemokines (Adachi et al., 1997; Matsukura et al., 1996), while viral proteins counteract this innate immune response. One example of a viral protein directly interfering on the protein level with cellular immune pathways is NS1 (its involvement in immune response suppression is reviewed in Hale et al. (2008)), which inhibits double-stranded-RNA-activated antiviral protein kinase (PKR). It is interesting to note that the lung subnetwork was also enriched for the "TLR JNK", "TRAF6 mediated IRF7 activation in TLR7/8 or 9 signalling", "IL-1 JNK", "TLR ECSIT MEKK1 JNK", and "IL1-mediated signaling events" pathways, because none of these are known to be specifically perturbed by viral proteins. IRAK1 (IL-1 receptor-associated kinase 1), which plays a critical role in IL-1 signaling events and in the activation of toll-like receptor (TLR) 7 and 9 pathways (reviewed in Gottipati et al. (2008)), is shared by all of these pathways. In non-stimulated cells, IRAK1 is associated with toll-interacting protein (TOLLIP), which is also part of the pathways listed above (with the exception of the "TRAF6 mediated IRF7 activation in TLR7/8 or

signalling" pathway). Stimulation of the IL-1 or TLR receptors leads to MyD88 (found in all pathways listed above) recruitment to the receptors. Through its interaction with MyD88, IRAK1 is also recruited to this complex. Complex formation results in IRAK1 phosphorylation and the activation of its kinase activity. Activated IRAK1 phosphorylates IRF7 or associates with TRAF6, resulting in IFNα induction and NFκB/MAPK activation, respectively. IRAK1 thus plays a critical role in innate immune responses, and the enrichment of IRAK1-dependent pathways in the lung network may contribute to the regulation of lung pathology in influenza virus infection.

A recent study demonstrated that signaling through the IL-1 receptor has a protective effect in mice infected with the pandemic 1918 influenza virus (Belisle et al., 2010). Another study reported that IL-1 receptor-deficient mice succumbed more easily than wild-type mice to infection with an H5N1 virus of low pathogenicity (A/Hong Kong/486/1997) (Szretter et al., 2007). Moreover, IL-1 receptor-deficient mice showed reduced inflammatory pathology upon infection with A/Puerto Rico/8/34 (H1N1) influenza virus (Schmitz et al., 2005). Several studies also established that influenza virus infection is sensed by TLR7 in plasmacytoid dendritic cell (Diebold et al., 2004; Geeraedts et al., 2008; Liang et al., 2011; Lund et al., 2004; Miettinen et al., 2001; Xing et al., 2011). However, none of these studies addressed the significance of IRAK1 in influenza virus pathogenicity. Our study thus exemplifies how our network analysis can identify potential regulators of influenza pathogenicity for experimental testing, for example, by assessing influenza virus infections in IRAK1-deficient cells or mice.

Next, we aimed to predict more specific novel interference mechanisms by constructing directed and tissue-specific protein networks linking the viral proteins with proteins whose corresponding transcript was up-regulated after influenza virus infection. We selected steadily up-regulated transcripts from a microarray experiment measuring gene expression changes over time in a lung epithelial cell line infected with a 2009 pandemic H1N1 virus (Li et al., 2011). To select steadily up-regulated genes, we filtered for probes differentially expressed at the last three time-points in the time series (30, 36 and 48h) with a q-value lower than 0.01 and a log2 fold change greater than 1. We selected 228 up-regulated transcripts in total.

As expected, all ten most strongly enriched pathways among the selected transcripts were involved in infection and the immune response. For example, most highly overrepresented was interferon alpha-beta signaling ($p < 10^{-20}$).

We constructed BET- and lung-specific networks connecting the viral proteins with the 228 up-regulated factors by shortest paths. From the shortest paths we assigned directions to edges on these paths (as described in section 4.2). The directed networks

consisted of 577 (BET) and 1056 (lung) PPIs. To examine if these networks might host relevant information on how viral proteins interfere with the cellular immune response, we tested for pathway enrichment in the reduced networks. We found the directed networks strongly enriched in immune response-related pathways (especially cytokine-related) even after excluding the 228 up-regulated transcripts, indicating that enrichment was independent of the high fraction of immune response factors in the transcriptomics data. For example, we observed a significant enrichment in both the reduced BET- and lung-specific networks for proteins related to IL-2 and IL-6 signaling and focal adhesions ($q < 0.05$). This suggested that directed and tissue-filtered PPI networks, indeed, might have captured relevant crosstalk between the viral proteins and immune pathways.

To mine the directed networks for interactions that are involved in interference mechanisms of the viral proteins with the cellular immune response, we concentrated, again, on layer one and two host factor proteins on the shortest paths. From the list of curated pathways enriched in both the BET and the lung directed networks, we selected several cytokine-related pathways and filtered for interactions where the second layer protein was in one of these pathways but the layer one protein was not (to specifically detect novel, indirect interference mechanisms). This resulted in a comprehensive BET network consisting of 49 interactions and a lung network formed by 67 interactions including viral proteins and host factors up to the second layer (see Appendix, Table 2).

Close inspection of these comprehensive cytokine-related networks in both BET and lung revealed several points of potential viral protein-mediated interference with inflammatory pathways. For example, both networks showed interactions between viral polymerase complex proteins (i.e., PB1 and PB2) and BHLHE40, a transcriptional regulator that is known to cooperate with HDAC1 to repress STAT1 activity (Ivanov et al., 2006). STAT1 is essential for the activation of interferon stimulated genes, which repress viral replication, and while influenza virus has an established ability to impair STAT1 (Pauli et al., 2008), no such function has been assigned to any of the viral polymerase complex subunits. In both comprehensive networks, BHLHE40 interacts with TOLLIP, a suppressor of TLR signaling (Cario and Podolsky, 2005) (see also the discussion of lung-specific inflammatory pathways above). This implies that the BHLHE40 protein could act as an important access point for influenza virus-mediated interference with host antiviral and inflammatory regulation, and further that viral polymerase subunits may have an important - yet unappreciated - role in this activity.

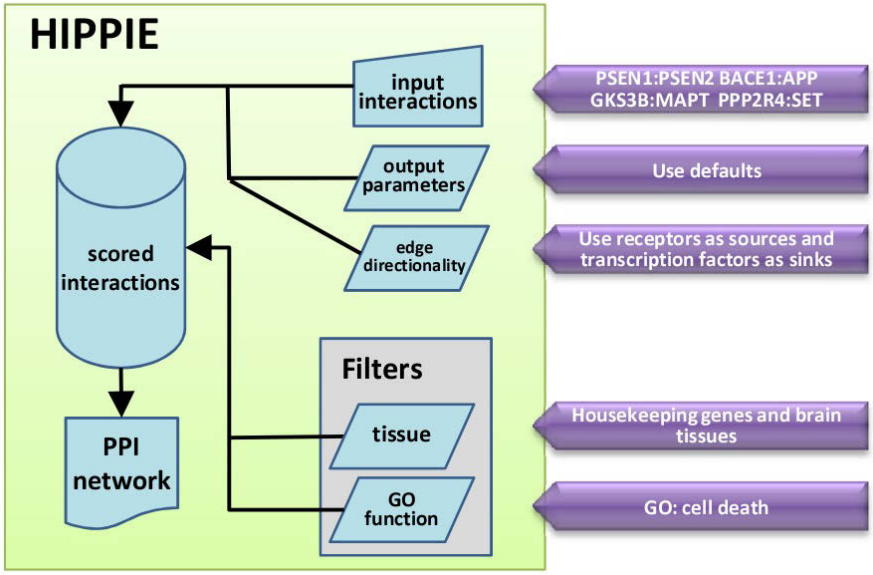

Figure 4.5: Protocol for the generation of a PPI subnetwork related to phosphorylation in Alzheimer's disease. The flowchart illustrates the input terms and options used to generate the network (see main text for details).

4.4 Search for phosphorylation-dependent protein-protein interactions related to Alzheimer's

Assuming no prior expert knowledge on a given topic, we applied a systematic protocol which can, in principle, be used to interrogate the PPI network about the involvement of protein interactions in a complex biological question according to current knowledge. In general, altered states of protein phosphorylation affect the PPI network and can lead to pathogenesis. Our goal in this example was to investigate the possible role of protein phosphorylation in Alzheimer's disease, the most common form of dementia. Alzheimer's disease is a degenerative disease manifesting in the brain, and its cause has been hypothesized to be the formation of protein aggregates leading to neuron death, in particular related to the abnormal phosphorylation of the microtubule-associated protein tau (Chun and Johnson, 2007).

To generate a list of PPIs related to Alzheimer's and protein phosphorylation, first, we used the webserver MedlineRanker (Fontaine et al., 2009) to retrieve a list of ranked PubMed abstracts (corresponding to manuscripts published within the last 5 years) according to their relevance to the search term "Alzheimer phosphorylation", which relates loosely to the question of interest. Next, we input the top 50 abstracts from

MedlineRanker into the webserver PESCADOR (Barbosa-Silva et al., 2011), which extracts a network of potential PPIs based on a set of PubMed abstracts. In our example, PESCADOR outputs 10 interaction pairs (type 2; co-occurrence of genes or proteins within a sentence containing a biointeraction term), of which only 4 pairs existed in HIPPIE as scored interactions (PSEN1:PSEN2, GSK3B:MAPT, APP:BACE1, PPP2R4:SET). We then studied the network surrounding these interactions.

The initial PPI network contained 726 interactions. Interactions could be further filtered on the basis of reasonable criteria (Figure 4.5), namely by tissue filtering for housekeeping and genes expressed in brain tissues (we selected "whole brain" and "prefrontal cortex"), and filtering for genes related to the GO term "cell death", reflecting that Alzheimer's disease is characterized by death of neural cells. Finally, to reveal potential signal transduction pathways we used the inference of edge directionality from receptors to transcription factors described above.

Within the resulting network, we highlighted the following path (Figure 4.6): LRP6-GSK3B-MAPT-AATF. The low density lipoprotein receptor-related protein 6 (LRP6) interacts with glycogen synthase kinase 3B and attenuates the kinase's ability to phosphorylate microtubule associated protein tau (MAPT) (Mi et al., 2006). Tau protein can contribute to Alzheimer's disease in different ways: 1) the hyperphosphorylation of tau protein can affect microtubule stability, leading to a disassociation of tau protein from the microtubule, possibly followed by the aggregation of phosphorylated tau into neurofibrillary tangles, which are observed in the brains of Alzheimer's disease patients (Dolan and Johnson, 2010); 2) mediated by protein phosphatase 1 and GSK3 activity, Tau filaments interfere with axonal transport in the neuron, which is consistent with deficiencies in axonal transport in Alzheimer's disease (LaPointe et al., 2009). Tau protein has been found to co-localize in the cytoplasm with Che-1 (AATF), which is an evolutionarily conserved RNA polymerase II binding protein that accumulates in the cell upon DNA damage (Claudio and Maurizio, 2007). It appears that Che-1/Tau proteins dissociate during neuronal cell death (Barbato et al., 2003); however, the function of Che-1 in the cytoplasm is unclear, as Che-1 is a nuclear protein that is involved in gene regulation of E2F1 targets and TP53 and has pro-proliferative and anti-apoptotic functions (Bruno et al., 2010). Together, these interactions suggest a complex interplay whereby the Tau phosphorylation state and structure, and context-dependent protein distribution within the cell may contribute to neuronal cell death and Alzheimer's disease pathology. An unbiased search for protein phosphorylation in relation to cell death in Alzheimer's disease pointed us to this interesting pathway.

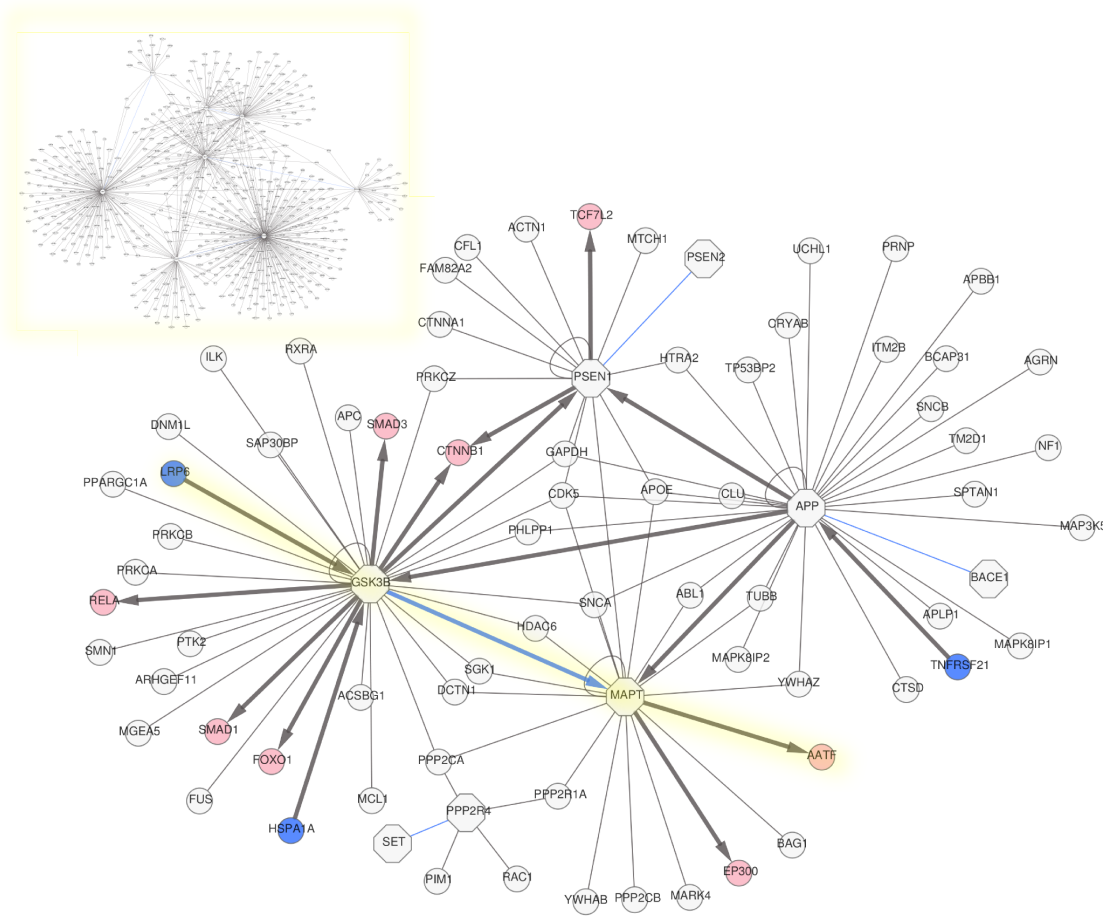

Figure 4.6: Generated PPI subnetwork related to phosphorylation in Alzheimer's disease. The network of the input interactions with their first neighbors is shown. The input interactions are are displayed in blue and the nodes they connect as octagons. Nodes corresponding to receptors and transcription factors are colored (blue and pink nodes, respectively). Edge directed path analysis from receptors to transcription factors resulted in the association of directionality to some of the edges (arrows). The path LRP6-GSK3B-MAPT-AATF is highlighted in yellow and described in the text. The inset shows the non-filtered network for comparison (726 interactions).

4.5 Discussion

The incorporation of tissue-specific expression information to create PPI subnetworks is a useful method to elucidate biological processes that cannot be observed when using the complete PPI network. Here we have shown an approach for the inference of associated context for PPIs based on the annotations of the interacting partners, which enhances the relevance of the annotated interactions. Interactions between proteins expressed in the same location (e.g., lung) or at the same time or developmental stage (e.g., embryo development) can then be selected. Directed pathways can be inferred and highlighted in the filtered network according to sets of sources and sinks corresponding to receptors and transcription factors. Using this approach we were able to identify novel, tissue-specific interactions between influenza virus proteins and cellular inflammatory signaling pathways that may regulate pathogenesis associated with infection, and to describe a brain-specific protein phosphorylation pathway relevant for Alzheimer's disease.

Several methods exist to create subnetworks of the human interactome based on context criteria. For example, POINeT (Lee et al., 2009) integrates the major PPI databases and allows the creation of tissue-specific networks. To our knowledge we are the first to combine edge directionality, gene expression and functional information for the detection of meaningful interactions. Some approaches exist that infer information flow in a network from the shortest path (or lowest costs if costs are associated with edges) that connects a set of source nodes with sink nodes. Cytoscape plug-ins such as BisoGenet (Martin et al., 2010) and GenePro (Vlasblom et al., 2006) find the shortest paths between nodes of the gene and protein network and represent properties of the nodes. SPIKE (Elkon et al., 2008) includes curated pathway data and also calculates pathway inference. The task of identifying signaling events from PPI data and functional protein annotation alone has been addressed in several studies (Mah et al., 2011; Vinayagam et al., 2011; Yosef et al., 2009) and implemented in tools (e.g., ANAT (Yosef et al., 2011)). Here, we proposed a protocol for edge directionality prediction based on calculating the shortest paths between sources and sinks. This protocol is runtime-efficient, which allowed us to integrate it with the web tool HIPPIE described in Chapter 3.

In summary, we have presented and made available an approach to associate context to PPI networks, which provides novel biological insight into mechanisms of disease. The continuing generation of PPI data and further incorporation into databases, and an increasing quality of annotations attached to genes and proteins will result in further improvements of our methodology.

4.6 Contributions

This chapter is a modified version of Schaefer et al. (2013). This work is based on ideas that we developed earlier and published in Lopes et al. (2011). In this earlier study we showed that disease-relevant pathways are more strongly enriched among specifically expressed (in cell types affected by the virus) interaction partners of viral proteins as compared to all interaction partners of the viral proteins.

All computational analyses described in this chapter were done by me except for the determination of the lowest common ancestors in the MeSH and GO hierarchies (done by Caroline Louis-Jeune and Carol Perez-Iratxeta) and the selection of genes specifically expressed in tissues (done by Tiago J.S. Lopes). The biological interpretation was done jointly with our collaboration partners at the Systems Biology Institute in Tokyo and at the University of Wisconsin-Madison. The paper was written by Miguel Andrade and me.

5 Evolution and function of polyglutamine in protein-protein interaction networks

5.1 Motivation

In this chapter, we describe how we investigated the evolution and function of polyQ stretches, which are known to cause neurodegenerative diseases when they undergo length expansion. For example, HD is caused by a length expansion of the polyQ stretch in the human protein huntingtin, which under non-pathological conditions ranges from 11 to 35 amino acids, to a length of over 40 residues. This drives the assembly of insoluble protein aggregates containing huntingtin together with other proteins. The length expansion of polyQ leads to cell death in the striatum and other regions of the brain, causing changes in movement and cognition (Reiner et al., 1988).

Besides their association with disease development, polyQ sequences may have an unknown physiological role (von Mikecz, 2009). While it was hypothesized that polyQ sequences might form a flexible spacer between protein domains like other low complexity regions (Faux et al., 2005; Huntley and Golding, 2000; Karlin and Burge, 1996), they are present in more than 60 human proteins (Butland et al., 2007), and some lower organisms, such as the amoeba *Dictyostelium discoideum* (Eichinger et al., 2005), possess several hundred. Anecdotal experimental evidence suggests a role of polyQ tracts in activation of gene transcription (Mitchell and Tjian, 1989). Accordingly, statistical studies revealed that proteins containing a polyQ stretch are biased toward functions related to transcriptional regulation and nuclear localization in several species (Alba and Guigo, 2004; Harrison, 2006; Karlin and Burge, 1996). Also, a more general role in mediating protein-protein interactions has been suggested (Hands et al., 2008).

In order to advance our understanding of the functions of polyQ regions in proteins, we investigated their potential properties from a systemic point of view. Since polyQ repeats have functional and evolutionary features that have been proposed to be relevant at different inter-related molecular levels, we have studied and combined analyses at the level of nucleotide sequences, protein sequences, protein structures and protein interac-

tion networks to obtain a systematic overview of polyQ function and evolution. We start with the analysis of CAG repeats at the nucleotide level, moving on to the analysis of protein sequences with polyQ tracts and phylogenetic studies of protein families, to the investigation of protein interaction networks of polyQ-containing proteins.

Our analyses suggest that the normal function of polyQ regions in proteins is to stabilize PPIs. We provide evidence for this hypothesis by analyzing the PPI network described in Chapter 3. We take into account the previously described study bias, which has a strong impact on network topological properties of well-studied proteins such as polyQ-containing proteins. As in Chapter 4, we will make use of functional protein annotations to generate biological hypotheses from network data. Here, we develop a strategy to control for the high amount of transcription factors among polyQ proteins when studying their associations to other proteins.

5.2 Distribution of CAG repeats in the human genome

Glutamines in proteins can be encoded by either CAG or CAA codons. However, the polyQ stretches that are enlarged in human disease proteins are encoded almost exclusively by pure CAG runs, while CAA repeats have not been observed (Gatchel and Zoghbi, 2005). This led to the suggestion of a possible mechanism for their generation by DNA slippage and hairpin formation during DNA replication, facilitating length extensions to which CAG but not CAA repeats are prone (Strand et al., 1993). This raises the question if CAG repeats, and the polyQ encoded by them, are just artifacts of faulty DNA replication without biological function.

This question can be answered by examining the genomic location of CAG repeats. If their genomic location was solely determined by random processes such as copy errors during DNA replication, and they underwent no evolutionary selection, they should be evenly distributed in the genome. On the contrary, if they had a biological function, their distribution should correspond to the molecular level of action: a function on RNA level would bias their genomic position toward transcribed genomic regions whereas a function in proteins would shift their distribution toward protein-coding exons.

We studied the distribution of CAG repeats of 10 or more consecutive trinucleotides in the human genome (GRCh37/hg19; Karolchik et al. (2003)). To measure whether the observed distribution is random or biased toward specific elements, we calculated the relative number of repeats located in different regions such as protein-coding exons, introns, untranslated regions (UTRs) and intergenic regions (conflicting assignments for a genomic region due to different splicing forms were resolved by giving priority

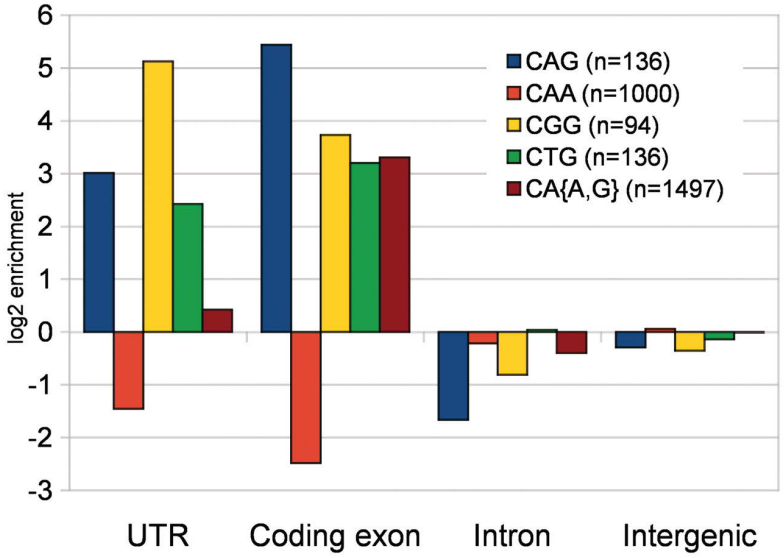

Figure 5.1: Frequency of trinucleotide repeats in the human genome. The y-axis represents the log2 of the ratio between the relative number of repeat runs observed (considering runs of at least 10 consecutive trinucleotides) and the proportion of the genome that is covered by the respective genomic region type.

to assignments in the following order: protein-coding exon, UTR, intron, intergenic region). These frequencies were normalized by the fraction of the genome covered by the respective region type (Figure 5.1).

Indeed, of 136 CAG repeats considered, 33 are in protein coding exons, resulting in a 43-fold enrichment over a random expectation, as previously described (Kozlowski et al., 2010). Although those 33 CAG repeats in coding regions could potentially encode three types of amino acid repeats depending on the reading frame (polyQ, polyS and polyA, for codons CAG, AGC and GCA, respectively), 28 coded for polyQ. This suggests that even if CAG repeats were accidental, they are selected for the encoding of polyQ in proteins, suggesting that polyQ has a biological function. Accordingly, the number of CAG repeats in introns and intergenic regions is close to random expectation (8 and 89, respectively; Figure 5.1). However, 6 CAG repeats are in UTRs (8-fold over random expectation), including the known disease locus in the 5'-UTR of the gene PPP2R2B causing spinocerebellar ataxia type 12 (Holmes et al., 2001), suggesting that they have a function at the transcript level. We also found CAG repeats enriched in UTRs and protein coding exons in rat (Baylor 3.4/rn4), mouse (NCBI37/mm9) and fly (BDGP R5/dm3), though to a lower degree than in human (UTR enrichment ranges from 1.7- to 2.5-fold and exon enrichment from 3.1- to 5.3-fold).

For comparison, we did an analysis considering consecutive runs composed of both codons encoding glutamine (CAG or CAA). These mixed trinucleotide repeats are 11 times more frequent in the human genome than pure CAG repeats. Like CAG repeats, the mixed repeats were enriched in exons and randomly distributed outside transcripts; their presence in UTRs, unlike CAG repeats, was close to random expectation. Together, these results suggest that CAG repeats have a function both at the protein and the transcript level.

We also analyzed the frequencies of pure CAA repeats in the different genomic region types. We found them to be generally more frequent in the human genome as compared to pure CAG repeats (1000 versus 136) but largely absent from protein coding regions (just one CAA repeat is located in a translated region encoding a polyQ stretch in the human protein ZFHX3). We note that there is a 2-fold enrichment of genes encoding tRNAs with an anticodon for CAG as compared to tRNAs matching CAA (21 versus 11) and that the CAG codon abundance is almost 3-fold higher in human exons (Chan and Lowe, 2009), but these numbers alone do not explain the 243-fold higher relative amount of CAG repeats in human protein coding regions.

Similarly, we did a calculation for CTG repeats in the human genome, which, like CAG repeats, are CG rich and when expanded are known to cause diseases of altered RNA function (Gatchel and Zoghbi, 2005) such as Myotonic dystrophy type 1 (Miller et al., 2000). Of a total of 136 CTG repeats, 7 were found in coding regions: 4 encoding for polyL (CTG codon), 3 for polyA (GCT codon) and none for polyC (TGC codon). As for CAG, the number of CTG repeats found in UTRs was significantly above random expectation, suggesting that they also have a biological function in transcripts.

As CGG repeats have been previously described as being the most strongly overrepresented trinucleotide repeats in human exons (Kozlowski et al., 2010), we also compared their distribution in the human genome to that of CAG repeats. We observed a similar distribution as the one for CAG and CTG with strong enrichment in UTRs and protein coding exons. In summary, whereas CAG repeats are clearly selected because they code polyQ in human proteins, there is some evidence of their function in non-coding parts of transcripts. We observed this in other mammals and for other CG rich repeats expanded in disease such as CTG.

5.3 Evolution of polyQ

After observing evidence for a function of CAG/glutamine repeats on protein level, we next analyzed the phylogenetic distribution of polyQ in proteins. We determined

the frequency of polyQ-containing proteins in a large number of species belonging to a wide taxonomic range (Magrane and UniProt Consortium, 2011). For this analysis (and hereafter, unless otherwise indicated), we identified as polyQ proteins those containing at least one polyQ stretch with a minimum length of 10 glutamines allowing for one mismatch (independent of its position within the polyQ tract) taking into account that polyQ stretches are often interspersed with single amino acids. This is, for example, the case in the known polyQ *D. melanogaster* Homeobox protein Deformed (positions 460 to 476: QQQAQQQQQSQQQQTQQ). The length threshold of 10 was chosen in order to account for all nine known human polyQ disease proteins (Gatchel and Zoghbi, 2005). Ataxin-7 is, among those disease proteins, the one with the minimum polyQ length of 10 residues in its non-expanded form. Using this definition, we found 86 human proteins with a polyQ stretch in the manually curated subset of UniProt (Swiss-Prot) (Magrane and UniProt Consortium, 2011).

We observed that the fraction of proteins having a polyQ stretch deviates greatly among different species being practically absent from prokaryotes. Figure 5.2 displays polyQ frequencies in several representative species from Swiss-Prot. This suggests that the abundance of polyQ proteins is not a random feature but depends on properties that vary between species.

To extend our analysis to an even broader taxonomic range we also calculated polyQ frequencies in the entire UniProt database (which includes the automatically curated sequences from TrEMBL). While the proteomes of bacteria and archaea typically contain no proteins with polyQ tracts at all, lower and higher eukaryotes on an average have 0.1% proteins with polyQ tracts. The *H. sapiens* fraction of polyQ proteins is above the average (0.34%) but much lower than that of many other organisms such as the yeast *S. cerevisiae* (1.1%), the fly *D. melanogaster* (3.8%) or the slime mold *D. discoideum* (10.5%).

5.3.1 PolyQ proteins in different organisms

In many cases, taxonomically related species have similar content of polyQ proteins but this is by no means the rule. For example, one can observe extreme differences between yeasts: the fission yeast *Schizosaccharomyces pombe* has only three polyQ proteins (out of 4974, $< 0.1\%$) whereas the baker's yeast *S. cerevisiae* has 79 (out of 6552, 1.1%), with other yeasts having even higher frequencies, e.g., *Neurospora crassa* (2.7%) and *Lodderomyces elongisporus* (6.8%). Variation of polyQ protein content can be significant even within species of the same genus. For example, in the 12 *Drosophilae* species that were analyzed the fraction ranges from 2.7% in *D. simulans* to the 8.9% of *D. grimshawi*

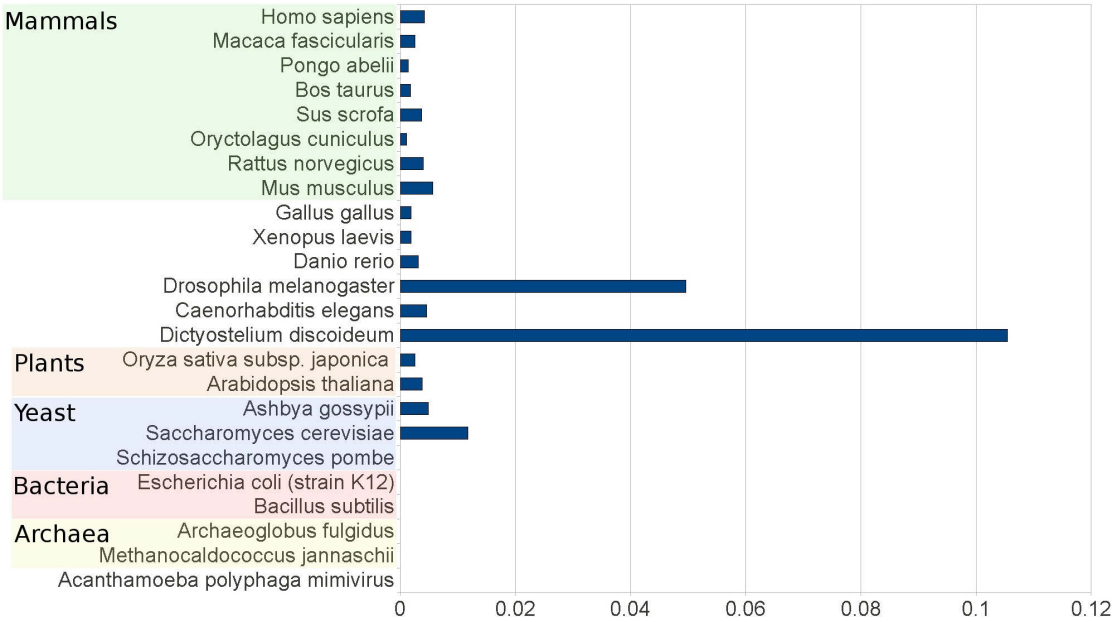

Figure 5.2: Relative amount of polyQ proteins in a representative set of species. The graph represents the fraction of proteins of each species' available proteome that contains a polyQ tract. Species with more than 1000 proteins in Swiss-Prot were included. Only two bacterial species are shown.

(median 4.2%). For the full list of polyQ abundance in species see Schaefer et al. (2012b).

Variation of polyQ protein content between species is generally high but we observed that it is lower when we compare closely related species, suggesting that it is tied to evolution. For example, while the three strains of the yeast *Paracoccidioides brasiliensis* analyzed had slightly different numbers of proteins, their overall polyQ frequencies were found to be similar (around 1.1%).

To find out whether there are species-specific functions that associated with polyQ protein content, we studied the frequency of polyQ proteins in a species in relationship to the presence of other proteins with particular domains. Protein domains are good indicators of particular protein functions and subcellular locations. We calculated the correlation between the relative number of proteins containing a polyQ stretch and the relative number of proteins containing a given protein domain over species. For this investigation, we used the protein annotations stored in the Pfam database version 23.0 (Finn et al., 2010), which hosts numerous accurate annotations of domains known or predicted to be present in proteins.

We computed the correlations of 4088 domains found in human proteins over all bacterial and eukaryotic species with at least 5000 protein entries in Pfam (for a total of

428 species, 133 of them eukaryotic and 295 bacterial). Since polyQ proteins are almost absent from prokaryotes, many domains appeared to be correlated to polyQ protein frequency mainly because they were exclusive to eukarya. Therefore, we additionally computed the correlation on the eukaryotic subset and used this value as a second selection criterion. We found 31 domains with a (Spearman) correlation value over all species > 0.8 and a correlation value on eukaryotic species > 0.35 (Table 5.1) indicating that these domains are generally enriched in the proteomes of species with many polyQ proteins and are underrepresented in proteomes lacking polyQ.

Among the most highly correlated domains were the FYVE and the PX domains. Remarkably, they are the only domains known to bind phosphatidylinositol 3-phosphate (PI3P) (Stenmark et al., 2002). Version 6 of the SMART database of domain annotations (Letunic et al., 2009) indicates that these domains do not co-occur in any of the current set of annotated proteins. This suggests that the identification of these two domains is based on independent sequences. The functional implication is that there is a true association to polyQ proteins and, more precisely, that the presence of polyQ proteins in a species is likely to be connected with processes that use PI3P, possibly in relation to signaling and transport mechanisms in which this molecule is involved.

We can point to further striking functional and structural similarities between the 31 correlated domains supporting associations of polyQ proteins to particular functions. Besides FYVE and PX, another three domains have a function in the phosphatidylinositol (PI) signaling system: CRAL/TRIO, PH, and Phosphoinositide 3-kinase family accessory domain. Two domains are related to ubiquitin (UBR box and UBX). Finally, we also observed five domains that belong to the zinc finger domain class: FYVE, UBR box, Zinc-finger double-stranded RNA-binding, Zinc finger ZZ type, and HIT zinc finger. Together, these observations suggest that polyQ proteins seem to be present in high numbers in species rich in proteins with roles related to PI signaling and the protein degradation system.

5.3.2 PolyQ emergence in protein families

Human non-pathogenic huntingtin contains an N-terminal polyQ tract of variable length ranging from 11 to 34 glutamines (Gatchel and Zoghbi, 2005). Such N-terminal polyQ appreciably and progressively shortens in orthologs from species increasingly distant from human along the chordate lineage (Q10 in dog, Q7 in mouse, Q6 in opossum, Q4 in *Xenopus* and fish; Figure 5.3, left box). We noted that the *Drosophila* huntingtin protein in various Drosophilae does not contain any N-terminal polyQ stretch but has several in two other regions of the protein (e.g., *D. yakuba* GenPept ID:195503512, has a Q10

Domain	Class	Correlation on eukaryotic subset	Correlation on all species
Importin-beta N-terminal domain		0.53	0.82
CAP-Gly domain		0.522	0.813
Zinc-finger double-stranded RNA-binding	ZF	0.494	0.823
Putative zinc finger in N-recognin (UBR box)	UBX, ZF	0.493	0.807
FYVE zinc finger	PI, ZF	0.479	0.829
Ku70/Ku80 N-terminal alpha/beta domain		0.476	0.806
GNS1/SUR4 family		0.47	0.8
Exportin 1-like protein		0.464	0.812
PX domain	PI	0.447	0.826
Mitochondrial carrier protein		0.435	0.824
PH domain	PI	0.432	0.817
Phosphoinositide 3-kinase family, accessory domain (PIK domain)	PI	0.431	0.803
Vps4 C terminal oligomerisation domain		0.428	0.813
G-patch domain		0.423	0.82
Mitochondrial ribosomal protein L51 / S25 / CI-B8 domain		0.417	0.818
RhoGAP domain		0.408	0.801
TBC domain		0.4	0.809
Domain of unknown function (DUF3434)		0.397	0.806
LisH		0.396	0.815
CRAL/TRIO domain	PI	0.393	0.814
La domain		0.393	0.802
MIF4G domain		0.391	0.803
GINS complex subunit Sld5		0.389	0.803
Region in Clathrin and VPS		0.382	0.819
UBX domain	UBX	0.377	0.811
Calponin homology (CH) domain		0.375	0.819
Phosphotyrosyl phosphate activator (PTPA) protein		0.37	0.807
emp24/gp25L/p24 family/GOLD		0.37	0.812
HIT zinc finger	ZF	0.367	0.801
Zinc finger, ZZ type	ZF	0.362	0.804
Vps51/Vps67		0.352	0.801

Table 5.1: Correlation of domains to polyQ presence over species. Structural classes of the domains are indicated: zinc finger (ZF), ubiquitin (UBX) or phosphatidylinositol (PI). Spearman's rank correlation over 133 eukaryotic species and over the full set of 428 species with high coverage in Pfam are given.

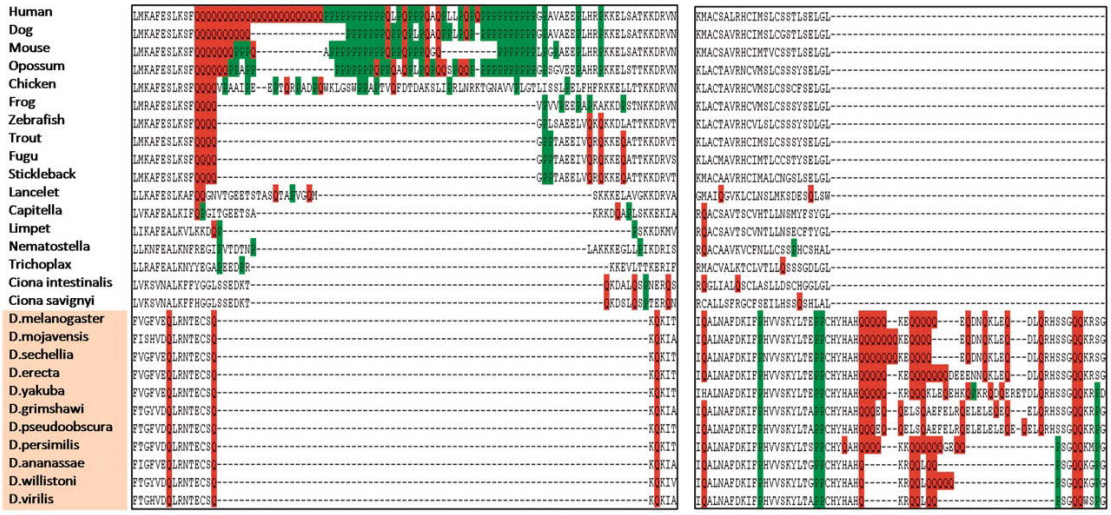

Figure 5.3: Fragments of a multiple sequence alignment of huntingtin orthologs from several species. Glutamines and prolines are marked in red and green, respectively. Left box: N-terminal polyQ region progressively enlarged along the chordate lineage and missing in Drosophilae. Note how this region is followed by polyproline in some species where the polyQ length is above four. Right box: very variable polyQ rich insertion specific to Drosophilae at another, distant position in huntingtin.

at positions 625-634 and a Q12 in a stretch of 14 amino acids at positions 1118-1131), which have no equivalent in the human protein (Figure 5.3, right box). This indicates that huntingtin proteins in ancestral species along the chordate and *Drosophilae* lineages have experienced independent events of insertion of polyQ tracts. This would suggest that the huntingtin protein is under evolutionary pressure to accept polyQ insertions, but that this pressure would not seem to act on the precise position of those insertions in the sequence.

To test whether this finding in huntingtin is unique or whether there are other protein families that underwent similar events during their evolution, we examined the distribution of polyQ-containing proteins in families of proteins with members in human (*H. sapiens*), zebrafish ((*Danio rerio*), representing another chordate) and the fly (*D. melanogaster*, representing a non-chordate organism). Given a protein family, existence of a polyQ in the human and fly proteins but not in the zebrafish protein will suggest that at least two independent events of polyQ insertion occurred: one outside the Chordate lineage and another within the chordate lineage, after the divergence of zebrafish and human.

We obtained 4759 protein families with at least one member from each of human, fly

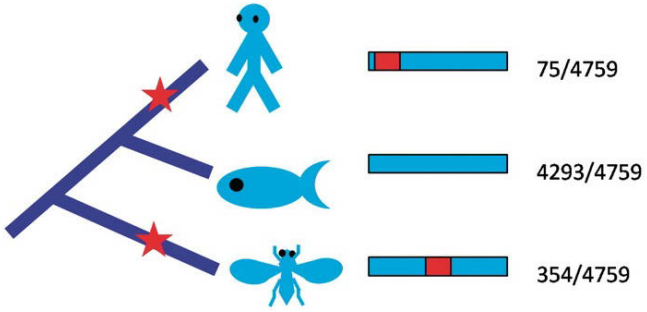

Figure 5.4: Protein families with multiple events of polyQ insertion. A total of 4759 protein families with members in human, zebrafish and fly was studied. We found 75 families having at least one human protein with a polyQ stretch, 354 families having at least one fly protein with a polyQ stretch, and 4293 having no Q-rich region in the fish proteins (see main text for details). For a total of 14 families (including huntingtin), both the human and the fly sequences had polyQ tracts (red boxes within the blue boxes) but not the zebrafish one, indicating multiple events of polyQ insertion along separate lineages (stars). By randomizing the identity of the polyQ sets in human and fly, we found the number of selected families to be significantly higher than random expectation ($p < 0.05$).

and fish according to the database of phylogenetic trees TreeFam (Ruan et al., 2008). We then selected those families in which the more distantly related species (human and fly) both had at least one homolog with a polyQ stretch (here requiring 8 Q in a range of 10 residues) while the zebrafish homologues were required to have no polyQ stretch at all (less than 5 Q in a window of size 10), considering them as families with evidence of multiple evolutionary events of polyQ insertion (see Figure 5.4). A total of 14 protein families fulfilled this conservative criterion. This number was significant ($p < 0.05$, randomization test). Considering also that in most cases the polyQ stretches appear at different protein positions within the aligned protein family, we conclude that the most likely explanation for the distribution of polyQ regions in the protein families analyzed is that polyQ emerged independently at different time points during evolution rather than that being lost in zebrafish.

The emergence of a significant number of protein families where insertion of polyQ tracts occurs in multiple ancestral proteins suggests that functional selection for the insertion of polyQ tracts at the protein level is a significant factor affecting the evolution of polyQ tracts. The fact that these insertions may be located at different positions in the protein suggests that polyQ performs a function that is not bound to a particular sequence. Insertion of polyQ tracts, however, does not seem to be absolutely necessary and therefore its function, while advantageous, must depend on some pre-existing, more

Category	HS	BT	RN	MM	DR	DM	CE	AT	SC	DD	NC
Transcription-related	✓		✓	✓	✓	✓		✓	✓	✓	✓
Nucleus	✓		✓	✓	✓	✓	✓	✓	✓		✓
(RNA and nitrogen) metabolic or biosynthetic process	✓		✓	✓		✓		✓	✓	✓	✓
Compositionally biased region (Ser,Gly,Pro,Ala)	✓	✓	✓	✓		✓			✓		
Protein phoshporylation	✓			✓		✓			✓	✓	
Alternative splicing	✓			✓		✓		✓			
Protein dimerization activity						✓		✓		✓	
Developmental protein						✓	✓	✓			

Table 5.2: Frequently overrepresented functional annotations among polyQ proteins from 11 eukaryotic species. HS = *H. sapiens*, BT = *B. taurus*, RN = *R. norvegicus*, MM = *M. musculus*, DR = *D. rerio*, DM = *D. melanogaster*, CE = *C. elegans*, AT = *A. thaliana*, SC = *S. cerevisiae*, DD = *D. discoideum*, NC = *N. crassa*. We merged the resulting species-specific lists of functional terms (applying a p-value threshold of 0.05 after multiple testing correction with the Benjamini-Hochberg method) and replaced similar terms by representative substitutes.

important functional context.

5.4 Protein context of polyQ

5.4.1 Function of polyQ proteins

It has already been noted that polyQ proteins are biased toward functions related to transcriptional regulation and nuclear localization (Alba and Guigo, 2004; Harrison, 2006; Karlin and Burge, 1996). To make a comprehensive analysis of the association of polyQ function to particular functions in the proteins containing it, we collected the polyQ proteins from 11 eukaryotic organisms of different taxa including plants, fungi, nematodes and chordates. We selected species with a sufficiently large coverage of the manually curated subset of UniProt (we required >750 protein entries in Swiss-Prot) and a minimum number of eight polyQ proteins to allow for conclusive enrichment statistics. From the resulting list we removed two of the three yeast members (*Kluyveromyces lactis* and *Candida albicans*) to end up with a diverse, representative set of eukaryotic species: *H. sapiens, B. taurus, R. norvegicus, M. musculus, D. rerio, D. melanogaster, C. elegans, Arabidopsis thaliana, S. cerevisiae, D. discoideum* and *N. crassa*. We then studied their functional annotations using two complementary approaches.

First, we computed the enrichment in GO annotations associated to these polyQ sets with respect to the total protein set. We analyzed each of the 11 species independently

using the web tool DAVID (Dennis et al., 2003). Annotations significantly enriched in the polyQ sets of several of these species included nuclear-related functions (e.g., transcription, splicing), but interestingly also protein dimerization, which correlates with our previously discussed findings suggesting that polyQ is involved in protein interactions (Table 5.2).

Second, we evaluated the enrichment of domains given by Pfam version 23.0; (Finn et al., 2010). A total of 14 domains were significantly overrepresented in polyQ proteins in at least 3 out of the 11 species (with corrected $p < 0.05$; see Table 5.3). However, we did not find these domains to be located closer than randomly expected to the polyQ or even overlapping the polyQ stretches.

Overall, most of these domains group into five functional categories: transcription regulation, protein binding, chromatin maintenance, RNA binding and signaling. For example, we found that more than a third of the domains enriched in the polyQ sets of at least three species are involved in protein-protein interactions (PAS fold, Bromodomain, PHD finger, PH, PDZ). Many of the associated domains fulfill functions in the nucleus.

In agreement with the association of polyQ protein content to PI signaling that we found at genomic level, we observed the PH domain, which is related to this function, to be enriched in the polyQ sets of three species. Many PH domains bind PI (10-20%), while others bind lipids, as well as peptides and proteins (DiNitto and Lambright, 2006).

5.4.2 Sequence features of polyQ flanking regions

The sequence environment of polyQ regions has been observed to influence the aggregation properties of polyQ-containing proteins, particularly polyproline (polyP) tracts (Bhattacharyya et al., 2006). In order to study the sequences surrounding polyQ for amino acid biases other than glutamine, we determined frequencies of amino acids in and around polyQ stretches in human proteins.

We found proline, histidine, alanine and methionine to be the four most strongly enriched amino acids around polyQ stretches in human proteins while cysteine, tryptophan, aspartic acid and isoleucine were the most under-represented (Figure 5.5). Analysis of other species revealed that the enrichment of proline and histidine in proximity of polyQ was conserved in *S. cerevisiae*, *D. melanogaster* and *D. discoideum*. We could not identify general amino acid properties (e.g., large size, high hydrophobicity) common to the amino acids in the under- and over-represented sets.

The described amino acid bias is not evenly distributed to both sides of the polyQ stretch. In human sequences the most extreme case is found for prolines, which often appear as polyP tracts almost exclusively C-terminal of the polyQ. For example, in the

Domain	HS	BT	RN	MM	DR	DM	CE	AT	SC	DD	NC
PHD-finger	✓		✓	✓	✓	✓				✓	✓
Bromodomain	✓		✓	✓	✓	✓				✓	✓
Helix-loop-helix DNA-binding domain	✓		✓	✓	✓	✓					✓
RNA recognition motif.				✓			✓	✓	✓	✓	
Homeobox domain	✓	✓	✓	✓		✓					
PAS fold			✓	✓	✓	✓				✓	
Helicase conserved C-terminal domain	✓						✓	✓		✓	
Protein tyrosine kinase						✓			✓	✓	✓
Protein kinase domain						✓			✓	✓	✓
PDZ domain	✓		✓			✓					
PH domain						✓			✓	✓	
Zinc finger, C2H2 type						✓	✓		✓		
ARID/BRIGHT DNA binding domain	✓			✓		✓					
SNF2 family N-terminal domain	✓						✓			✓	

Table 5.3: Domains overrepresented in polyQ proteins from eleven eukaryotic species. HS = *H. sapiens*, BT = *B. taurus*, RN = *R. norvegicus*, MM = *M. musculus*, DR = *D. rerio*, DM = *D. melanogaster*, CE = *C. elegans*, AT = *A. thaliana*, SC = *S. cerevisiae*, DD = *D. discoideum*, NC = *N. crassa*.

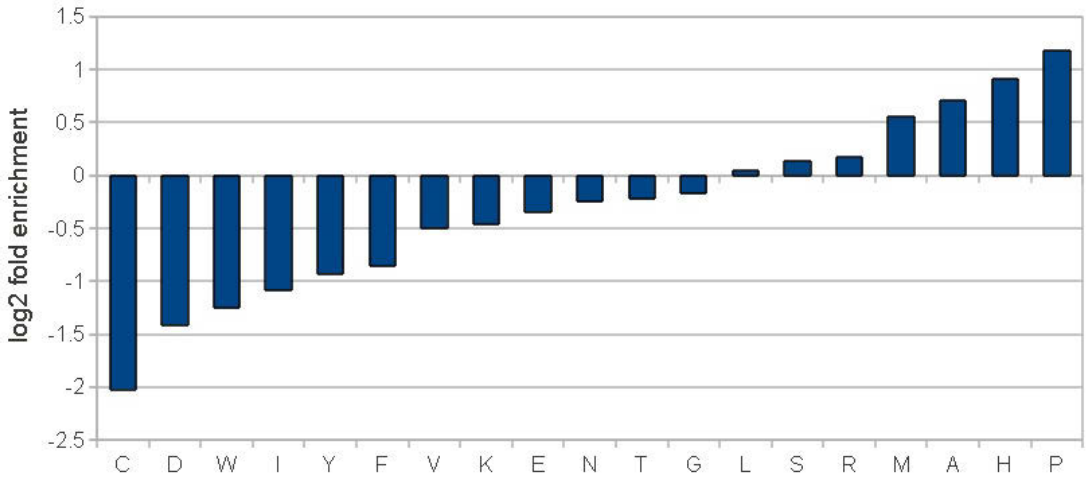

Figure 5.5: Amino acid usage in flanking sequences of polyQ. Amino acid usage in human proteins was calculated for polyQ stretches and for the 50 amino acids flanking polyQ. The y-axis represents the log of the ratio of the frequency of the amino acid in the flanking sequence to the frequency of the amino acid in the human genome.

set of 86 human polyQ proteins, 13 proteins contain a polyP run of at least three residues at a maximum distance of three amino acids from the polyQ. Of those we found twelve C-terminal and only one N-terminal to the polyQ stretch. This agrees with the findings of Bhattacharyya et al. (2006) that polyP stretches inhibit polyQ-dependent aggregate formation only when located C-terminal of the polyQ tract. The other enriched amino acids (alanine, methionine, histidine) tend to occur as single amino acid residues and the number of repeated amino acid tracts surrounding polyQ was too small to assess any distributional bias.

In a recent computational study polyQ tracts were associated with the presence of disordered regions (Simon and Hancock, 2009). To quantify this association, we applied the stand-alone version of the tool RONN (Yang et al., 2005) for the prediction of disordered regions. We considered residues with a probability above 0.8 to be disordered. We found 96 regions predicted as disordered at a distance of 10 or less residues from polyQ tracts in human proteins, which affects the vast majority of all 109 polyQ stretches. To test whether this finding is due to the polyQ tract itself being predicted as disordered, we repeated the computation, this time removing the corresponding polyQ tracts from the sequence. From the 96 regions, 35 remain disordered after polyQ tract removal. To assess the significance of this observation, we calculated estimates of the background frequency of disordered regions among all human proteins by randomly sampling 1000 human proteins from the Swiss-Prot database and determining the proportion of residues predicted to be disordered among all residues within the sample. The enrichment of disordered regions around polyQ deletion sites was significant ($p = 6.2e\text{-}11$).

5.5 PolyQ in PPI networks

5.5.1 PolyQ in protein complexes

Both, previous studies (Hands et al., 2008) as well as the above described functional association of the polyQ set to protein dimerization, indicate an involvement of polyQ in PPIs. To look for further evidence supporting this hypothesis, we investigated whether polyQ-containing proteins are enriched among proteins that form complexes. Among 1825 human protein complexes defined in Ruepp et al. (2008), we identified 130 having at least one protein containing a polyQ stretch (using the same polyQ definition as above: repeat length of 10 glutamines allowing for one mismatch).

These 1825 human complexes are formed by 8797 components; among them 149 are polyQ proteins, showing a 4-fold enrichment with respect to the frequency of polyQ proteins in the human proteome. In the non-redundant list of 2541 proteins participating

in complexes, the enrichment is still significant (2.1-fold). This suggests that polyQ proteins even have a tendency to form part of multiple complexes. Examples of human polyQ proteins that are members of several complexes are CBP and TBP.

To test whether there is a significant tendency to find multiple polyQ proteins within individual protein complexes, we applied a randomization test. We randomized the polyQ annotations and observed whether we obtained an equal or larger amount of complexes containing two or more polyQ proteins, which happened in 52 of 1000 tests ($p = 0.052$). For less restrictive polyQ threshold selections, the results were even more significant (e.g., eight Qs in a window of 10 residues resulted in $p < 0.001$). This suggested that polyQ-containing proteins are not randomly distributed among complexes but that the chance of seeing one polyQ protein increases significantly the chance of finding at least one other polyQ-containing protein in the same protein complex. For example, the RSmad complex contains a total of 10 proteins. Among them are three polyQ-containing proteins: ARID1B, CBP and NCOA3. In summary, protein complexes are enriched in polyQ proteins suggesting that polyQ function is related to protein interactions.

5.5.2 PolyQ tracts are associated to proteins with many partners

To further investigate the association of polyQ with protein interactions, we compared the distribution of polyQ-containing proteins and the number of protein interacting partners (according to the HIPPIE database of human PPI data described in Chapter 3). We observed that proteins containing polyQ have significantly more interactions than proteins that do not ($p < 10^{-9}$, Wilcoxon-Mann-Whitney test). However, we observed that polyQ proteins have a longer than average length (1253 residues versus 550 residues) and that longer proteins have more interaction partners as compared to short proteins, probably due to their higher number of potential interaction interfaces (e.g., the longest 25% of all human proteins have a mean value of 9.1 interaction partners while the shortest 25% have only 4.9). Therefore, we repeated the test comparing the polyQ set with proteins at least as long as the average polyQ-containing protein. The resulting p-value of 0.007 indicated a higher number of interaction partners for the polyQ set.

Since it is likely that transcription factors have more interactions than the average protein and polyQ proteins are enriched in transcription factors, we repeated the test comparing the interaction distribution of the polyQ proteins to that of the set of human transcription factors as defined by UniProt annotations (Magrane and UniProt Consortium, 2011) without a polyQ tract (Figure 5.6A). The resulting p-value of 0.009 was once again significant.

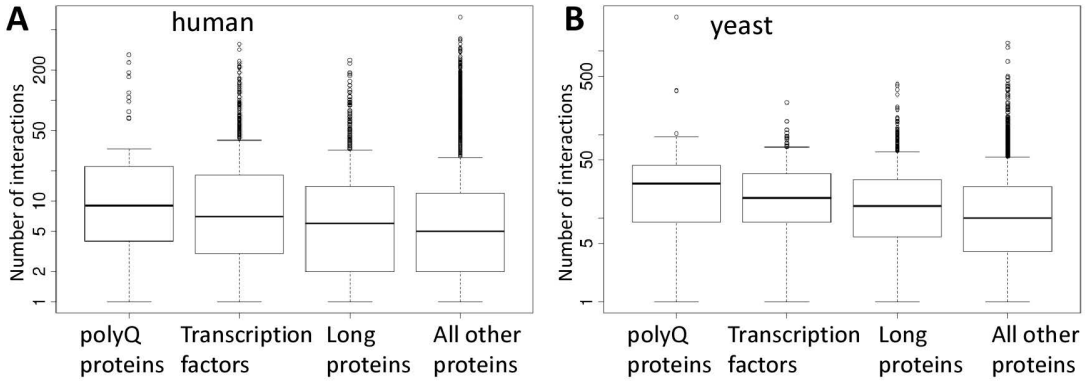

Figure 5.6: Protein interaction degree distribution for different classes of proteins. Box plots of the distribution of protein interaction partners for different protein sets. (A and B) Comparison of polyQ proteins, transcription factors without polyQ, large proteins without polyQ and all non-polyQ proteins, for human and yeast, respectively. All pairwise differences within a species were significant ($p < 0.01$).

As we have shown in Chapter 3, PPI networks are highly biased by the experimental selection of bait proteins. As a consequence, the measured number of interaction partners of a protein depends not only on its true physiological number of interactions but also on how intensively the protein has been studied in PPI assays. Since several polyQ proteins known as disease-causing agents have been studied multiple times, we assume that this could have an impact on the amount of reported interactions of polyQ proteins. Indeed, when comparing the bait usage numbers within the polyQ set with those in the entire proteomic background, we find a higher fraction of polyQ proteins that have been used several times as a bait (24.4%) than among the non-polyQ proteins (6.7%). To rule out the possibility that the observed higher number of interactions formed by polyQ proteins is solely due to the selection bias, we chose two strategies to control for the higher bait usage frequencies among polyQ proteins: (a) we repeated our analysis in another species in which we do not expect a study bias towards polyQ proteins and (b) compared the mean number of interaction partners of polyQ proteins to randomly sampled sets with the same bait usage distribution.

We carried out the analysis with the proteins of *S. cerevisiae* (BioGRID v3.1.74; Stark et al. (2011)) and observed that its polyQ proteins have a significantly higher number of interacting partners than those that do not have polyQ ($p < 10^{-12}$), even when filtering for transcription factors ($p = 0.041$) or proteins of higher length ($p = 0.0003$) (Figure 5.6B). These results confirm that our findings are not species-specific and suggest that our observations in human proteins are not due to a bias in the PPI network arising

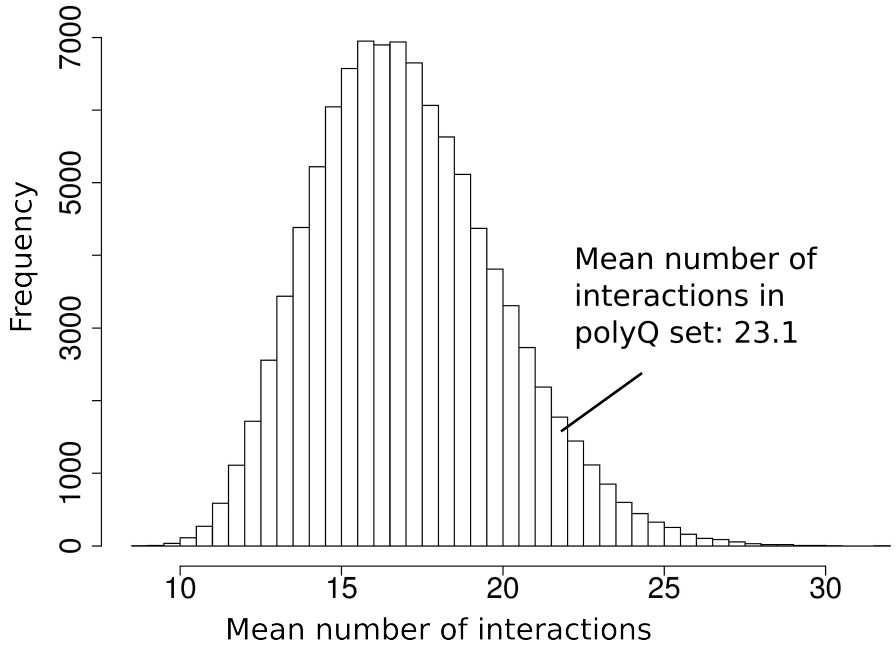

Figure 5.7: Mean interaction number of proteins with same bait usage distribution as polyQ set. We find the number of observed interactions of the polyQ set significantly larger.

from researchers focusing on particular human proteins related to disease.

To directly control for the high fraction of polyQ proteins that have been screened several times for interaction partners, we randomly generated sets of proteins where polyQ proteins that have been studied multiple times have been replaced by non-polyQ proteins which have been used the same number of times as a bait protein. For each of the 10,000 generated random sets, we calculated the mean number of interaction partners (Figure 5.7) and found the observed mean value for the polyQ set significantly larger ($p = 0.028$). We repeated the comparison requiring the randomly sampled proteins to be transcription factors and having a similar length as polyQ-containing proteins and found both distributions of mean number of interaction partners significantly lower than the observed value for the polyQ set ($p < 0.05$).

To test whether there is an effect of the length of the polyQ stretch on the number of interactions, we binned either all human or all yeast proteins into the three categories: lacking polyQ tract, having a small polyQ stretch (length between 5 and 14 amino acids), and having a long polyQ tract (longer than 14 amino acids). As in the analyses described earlier, we counted the number of interactors for each of these proteins (Figure 5.8). The differences between the degree distributions were significant ($p < 0.01$)

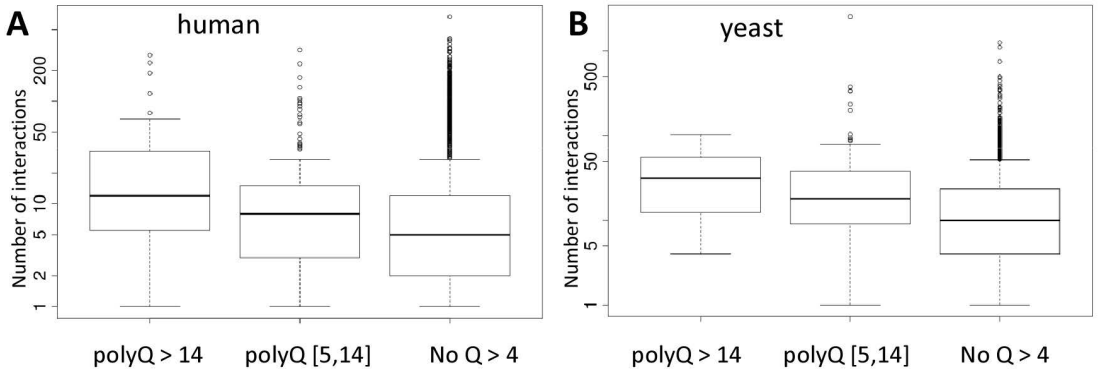

Figure 5.8: Protein interaction degree distribution for different lengths of polyQ. Box plots of the distribution of protein interaction partners for different protein sets. (A and B) Comparison of proteins with long polyQ, short polyQ, or no polyQ, for human and yeast, respectively. All pairwise differences within a species were significant ($p = 0.01$) except for the comparison between medium and long polyQ length in yeast ($p = 0.056$). This exception was due to an outlier in the medium set: one of the proteins has a degree of 2549 which is more than twice as high as the second highest degree. Removing it results in significant differences for all comparisons.

and increased with the length of the polyQ tract both for human and for yeast proteins. This observation suggests a correlation between the length of polyQ and the interaction capacity of the hosting protein.

In summary, these analyses demonstrate that polyQ proteins have more protein interactions than proteins lacking a polyQ tract. Although there is a component in that effect related to polyQ proteins having longer than average length, being associated to particular functions and being used as baits more often than other proteins, these properties of polyQ proteins alone are not responsible for the whole effect. We interpret these results as indicating that polyQ tracts favor PPIs.

5.5.3 Function of proteins interacting with polyQ proteins

We found that polyQ proteins are associated with interactions between proteins and that polyQ proteins are enriched in some general functions related to the nucleus such as transcriptional regulation and chromatin maintenance among others. We wondered whether there would be particular domains or functions specific to the proteins interacting with polyQ proteins. These would account for other indirect functional associations of polyQ proteins.

To investigate this, we measured the significance of over-representation of predicted domains in non-polyQ proteins that interact with polyQ proteins. To detect overrep-

resented domains in the set of proteins interacting with the polyQ proteins in the PPI network, a randomization test was applied. As a test statistic, the number of interactions between either the polyQ or the random set, and the domain set was calculated. The background distribution was generated by selecting a set of non-polyQ proteins, preserving the degree distribution and the fraction of transcription factors of the polyQ set.

In our analysis of proteins interacting with polyQ proteins, we found 17 domains significantly enriched (occurring in more than 10 polyQ-protein interacting proteins; $p < 0.01$). The list of domains was manually curated to remove redundant entries and obvious false positive predictions (as detailed in Schaefer et al. (2012b)). The set was different from the set of domains found to be enriched in polyQ proteins with the exception of nuclear hormone receptor domains (NHR and Zinc finger C4 type). In general functional terms, the list of domains once again contained an important fraction of domains with nuclear functions (NHR, bZIP, MH1, Zinc finger MIZ type). However, it also included domains with non-nuclear functions (Ubiquitin family, AAA, EGF). The curated set of domains enriched in proteins interacting with the polyQ set is listed in Table 5.4.

For comparison, we tested the enrichment of protein domains in proteins interacting with polyP tracts (defined as for polyQ tracts: minimum number of 10 consecutive P allowing for one mismatch). The known polyP interacting domains SH3 and WW were among the top 15 most enriched domains (additionally, Actin, RhoGEF, SH2, BAR, Spectrin, Arf, FF, PX, FCH, WH1, CH, RhoGAP and the UBX domain) all being significantly associated to polyP ($p < 0.01$). The strength of these associations is comparable to that of the top polyQ associated domains.

5.5.4 PolyQ as a motif for protein interaction

Multiple observations presented above seem to indicate that polyQ is involved in PPI: polyQ proteins are related to dimerization, proteins with longer polyQ tracts tend to have more interaction partners, and many human protein complexes contain multiple polyQ proteins.

In addition, among the 86 human polyQ-containing proteins, we counted 49 interactions where both interacting proteins contained polyQ tracts. Randomizing the identity of the polyQ set, we found this enrichment as significant as the enrichment determined for domains present in proteins interacting with polyQ proteins ($p = 0.0023$).

Moreover, among the list of domains enriched in proteins that interact with human polyQ proteins, we observed the bZIP domain, which can form a coiled-coil. It was

Domain	P-value	Number interactions
Nuclear hormone receptor associated	< 0.001	95
EGF	< 0.001	29
ATPase family associated with various cellular activities (AAA)	< 0.001	24
Zinc finger MIZ type	0.002	12
Ubiquitin family	0.002	25
MH1 domain	0.007	30
Basic Leucine Zipper Domain (bZIP)	0.009	37

Table 5.4: Domains overrepresented in proteins that interact with human polyQ proteins.

recently shown that polyQ regions overlap with coiled-coil regions in a set of polyQ-containing proteins and are also found in their interaction partners (Fiumara et al., 2010). Coiled-coil domains are involved in oligomerization. This would explain the function of polyQ observed above as a ubiquitously used motif of protein interaction.

To determine whether the association of polyQ tracts and coiled-coil regions is a general phenomenon, we systematically studied the overlap between polyQ regions and predicted coiled-coil regions in polyQ proteins. For the prediction of coiled-coils, we applied the tool Coils (Lupas et al., 1991), which detects hydrophobic heptad repeats in protein sequences. We considered only high-confidence predictions (over a probability threshold of 0.8). We observed a significant enrichment in human and in the other 10 eukaryotic species analyzed (we studied the same set of representative species as used in section 5.4.1). For example, of the 109 polyQ tracts in 86 human proteins, 54 (50%) overlapped with a coiled-coil region and 5 more were in very close proximity (distance of 10 amino acids or less) ($p < 10^{-15}$).

We found that the distribution of coiled-coils is extremely biased toward the N-terminus of the polyQ tract. In this respect, one has to note that the amino acid composition of the regions surrounding polyQ tracts is biased for some amino acids. As described in section 5.4.2, we could detect enrichment for several amino acids around polyQ tracts in several organisms and the described amino acid bias is not evenly distributed to both sides of the polyQ stretch. In human sequences, proline is most strongly enriched and often appears as a polyP tract C-terminally to the polyQ.

To exclude the possibility that the bias of coiled-coils toward the N-terminus of polyQ tracts is simply due to C-terminal polyP, we analyzed the position of coiled-coils with respect to polyQ tracts excluding cases where polyP was present. The bias was still

observed both in human (34 N-terminal versus 6 C-terminal) and yeast (14 N-terminal versus 1 C-terminal), suggesting that the association of coiled-coils to polyQ tracts is asymmetric.

To exclude the possibility that the observed colocalization of coiled-coils with polyQ stretches is an artifact of the prediction tool that was applied (e.g., over-predicting spurious coiled-coil regions on polyQ stretches), we repeated the coiled-coil prediction on the human polyQ set with a different coiled-coil prediction approach using the tool Paircoil2 (McDonnell et al., 2006). We found, again, a significant enrichment of coiled-coils in the polyQ set: for a tool specific threshold of 0.025, we observed 30 proteins with a coiled-coil among the 86 human polyQ proteins (34.9%) at a background prediction rate of 13% ($p < 10^{-9}$).

To further substantiate our observations, we deleted the polyQ stretches from the sequences and repeated the coiled-coil prediction in 45 human proteins hosting 54 polyQ stretches that were either overlapping a coiled-coil or in close proximity of one. We excluded from this analysis those proteins where the coiled-coil was predicted to be within a polyQ stretch and those with a C-terminal polyP. We counted how often we could still observe a coiled-coil prediction in the 10 residues flanking each side of the 54 deletion sites. In 11 of the 54 cases, a coiled-coil was predicted, which corresponds to a 6-fold enrichment over the background frequency of predicted coiled-coil regions in all human protein sequences. This enrichment was significant ($p < 10^{-6}$; probability of observing 11 or more coiled-coil regions under the Binomial distribution). This result proves that the association of polyQ to coiled-coil regions is not just due to the presence of polyQ but that also its flanking sequences have significant coiled-coil forming potential. In addition to the enrichment of coiled-coils in proteins with a polyQ stretch, we could also establish that non-polyQ proteins interacting with polyQ proteins are significantly enriched in coiled-coil regions ($p < 0.001$).

In summary, we found a significant association of polyQ tracts after coiled-coil regions. This agrees with a function of polyQ tracts related to protein interactions. We wondered if functions previously noted to be associated to polyQ proteins (Alba and Guigo, 2004; Harrison, 2006; Karlin and Burge, 1996) could be just a secondary effect and explained simply by the fact that those functions (e.g., transcriptional regulation) require more protein interactions than other functions (e.g., metabolism). Therefore, we tested if we observed a similarly high enrichment in certain GO terms when we compared proteins with many interactions to all proteins with at least one known interaction. Indeed, many functions associated to the polyQ set are also enriched in the set of proteins with the 10% highest number of interaction partners. For example, both in yeast and human,

we observed in the set of proteins with many partners a significant enrichment of the GO term GO:0031981-nuclear lumen (p-values of 6.3e-53 and 4.2e-28) and in human proteins of the term GO:0008134-transcription factor binding ($p < 10^{-23}$). This effect is independent of the precise protein set size and can be reproduced with the 86 highest degree proteins in human (a cutoff chosen in accordance with the size of the human polyQ set).

5.6 Discussion

PolyQ tracts in protein sequences have been researched mostly because of their pathogenic expansion in multiple human genetic diseases. However, their presence in many wild-type proteins across a variety of species is intriguing and suggests that normal polyQ tracts might have a function.

Following this idea, we provided evidence collected at multiple inter-related biological levels that collectively and consistently indicates that polyQ tracts are involved in protein interactions, e.g., because of their enrichment in protein complexes, and their association with coiled-coil regions. Through our analyses, we noted other features of polyQ tracts, which may not be directly related to their function as an interaction motif, but to the pathogenic effects of their abnormal expansion.

At the nucleotide level, we could observe selection of CAG repeats in exons of human, mouse, rat and fly genes. Intriguingly, we also found them enriched, although at a lower level, in UTRs. It was noted that both CAG and CTG repeats can form RNA-DNA hybrids (R loops) that could have a biological function (Lin et al., 2010; Reddy et al., 2011). In agreement with this, we found CTG and CGG repeats similarly enriched in UTRs but not so much in exons.

Along these lines, we found that whereas only 13% of prolines in human proteins are encoded by the rare codon CCG, this fraction is higher in prolines forming polyP (of length three or more) (23%), and even higher (43%; $n = 48$ codons) if the polyP is near uninterrupted polyQ sequences of minimum length 10 (at a maximum distance of three amino acids). We observed a related effect in polyQ, which are encoded more frequently by CAG codons when the polyQ sequences are close to polyP tracts ($n = 156$) as compared to other polyQ ($n = 1169$) (90% versus 79%). This inter-dependence between GC rich codons hints at an effect at the transcript level. In summary, we interpret these results as indicating that CAG repeats are under positive selection at the nucleotide level. Abnormally expanded CTG repeats bind muscleblind resulting in Myotonic dystrophy type 1 (Miller et al., 2000). CAG and CTG repeats might bind to

proteins in their wild-type transcripts.

The mechanisms that have been proposed to originate regions encoding poly-amino acid repeats are currently not well understood (Kovtun and McMurray, 2008). Many polyQ tracts are composed exclusively of CAG repeats in mammals (Alba et al., 2001) (the other possible codon encoding Q being CAA); this is interpreted as evidence of their formation due to trinucleotide expansion by gene slippage, resulting from the formation of an abnormal loop of the CAG repeat via CG pairings. According to this, it was shown that polyQ-coding pure CAG repeats are expanded from mouse to human (Hancock, 1995) while expansion does not occur if they are formed by a mix of CAG and CAA codons. In contrast, in some non-mammalian organisms, for example *Drosophila* (Alba et al., 2001) and *D. discoideum* (Eichinger et al., 2005), polyQ tracts tend to be encoded by pure CAA repeats actually suggesting that they are selected to resist slippage.

The abundance of proteins with polyQ tracts across different species is highly variable, for example they are completely absent from prokaryotic organisms. In a few species, polyQ tracts are among the most frequent amino acid repeats (Faux et al., 2005; Karlin et al., 2002). This variability may be related to the inability of some species to deal with these aggregation-prone repeats. Therefore, analysis of the correlation between systemic properties of species and presence of polyQ proteins might hint at the mechanisms by which species deal with polyQ proteins and at the origin of their pathogenic effects. We investigated this systematically and observed a huge variation between species in content of polyQ proteins (e.g., highest in *D. discoideum* and *D. melanogaster* but very low for *Xenopus* or *D. rerio*). We observed that those species with high polyQ protein content have a higher number of proteins bearing domains with functions related to PI signaling and ubiquitin-directed protein degradation. Interestingly, both ubiquitin and PI play a role in the clearance of polyQ aggregates: aggregates containing proteins with an expanded polyQ stretch have been shown to be ubiquitinated (Suhr et al., 2001a), while PI-binding domains are involved in targeting polyQ aggregates to membranes during the process of macroautophagy. For example, the FYVE domain containing human protein Alfy promotes the degradation of huntingtin in mammalian cells (Filimonenko et al., 2010). Therefore, this association between high content of polyQ proteins at the genomic level to both PI signaling and ubiquitin-directed protein degradation could be explained by the need of the cell to effectively degrade polyQ-containing aggregation-prone proteins. We speculate that differential selection explains the high content of polyQ proteins in some organisms (Hancock, 1995): organisms that can select polyQ co-evolve the appropriate machinery to clear polyQ protein aggregates, whereas organisms lacking strong clearance mechanisms for protein-aggregates might not tolerate polyQ

proteins at all; this could explain the absence of polyQ proteins in the prokaryotes.

When analyzing the variability of polyQ protein occurrence in a large variety of species in more detail, we observed that polyQ tracts are not a feature characteristic of particular gene families. Variability of polyQ protein content among species is therefore due to orthologs of a protein having a polyQ tract in one species and not in another. Moreover, we demonstrated that particular protein families show multiple events of emergence of polyQ and that they can happen at different positions of the sequence. For example, the human huntingtin has a polyQ tract situated near the N-terminus of the protein, whereas many lower organisms have none, e.g., *Ciona* (Gissi et al., 2006). However, the huntingtin of the *Drosophila* genus has multiple polyQ tracts, none of them in the N-terminus. This suggests that particular protein families, including huntingtin, are under selective pressure to accumulate polyQ tracts. In summary, the fact that some organisms have orthologs lacking the polyQ tract indicates that the protein can fulfill its tasks without this feature: its function cannot be essential. In addition, the fact that the polyQ tracts can occupy different positions in the sequence suggests that polyQ tracts perform a function without strong positional requirements. On the other hand, they do have some function specific to particular protein families since evolutionary pressure to insert the polyQ tracts leads to this occurring in distantly related clades. In terms of speed, the evolutionary expansion of a polyQ tract is much slower than a pathological expansion. For example, the expansion of the N-terminal polyQ tract in human huntingtin could be estimated to have evolved at an average rate of one Q per 30 million years. This is indicative of how delicate the effect of modification of polyQ tract length can be (Figure 5.3, left box). On the other hand, the large variations of the huntingtin mid-of-sequence polyQ tracts in the different *Drosophilae* indicate that fast evolution of polyQ tracts is also possible (Figure 5.3, right box).

There is already some experimental evidence suggesting that the function of polyQ could be to modulate PPIs. For example, a polyQ sequence in TBP modulates its interaction with TFIIB (Friedman et al., 2007), and a glutamine-rich activation domain in SP1 directly interacts with TAF4 in *Drosophila* (Hoey et al., 1993). It was observed in an *in vitro* experiment that mouse Sp1 and some components of the core transcription apparatus (e.g., TFIID and TFIIF) are direct targets inhibited by mutant huntingtin in a polyQ-dependent manner (Zhai et al., 2005). In addition, mutant proteins with enlarged polyQ tracts aggregate, which also points to a relation between polyQ and protein interactions (Kopito, 2000; Ross, 1997; Tran and Miller, 1999).

We were able to provide further evidence at multiple levels to support the hypothesis that polyQ modulates PPIs. For example, we could detect that polyQ proteins have

more protein interaction partners than non-polyQ proteins and have a higher tendency to interact with other polyQ proteins than non-polyQ proteins.

In agreement to previous studies (Whan et al., 2010), we identified an over-representation of protein domains related to nucleus-based functions in polyQ proteins but also in proteins that interact with them. In fact, we could demonstrate that these functions are also over-represented in proteins with many interactions. Therefore, we deduce that the functional biases observed in polyQ and polyQ-interacting proteins are due to the involvement of polyQ in protein interactions.

Until now, the structural basis for the possible modulation of protein interactions by polyQ is not clear. To begin with, the precise structure of polyQ itself is unknown and suggested conformations of both synthetic polyQ peptides and naturally occurring proteins with polyQ tracts include alpha helix, random coil, and extended loop as it has been described for huntingtin exon1 (Kim et al., 2009; Li et al., 2007). This might be due to polyQ adopting an unstable context-dependent structure. Part of this context can be flanking sequences, which have been shown to influence both the structure (Kim et al., 2009) and aggregation properties of polypeptides with polyQ tracts (Dehay and Bertolotti, 2006). Length expansions of polyQ stretches seem to be accompanied by a transition of a random coil into a beta sheet structure (Perutz, 1996), which would account for its pathogenic effect. In addition, polyQ tracts seem to be able to modify the conformation of structured domains nearby in sequence (Ignatova and Gierasch, 2006). Such interactions could be dependent on the presence of other interacting proteins, and it has recently been suggested that the mechanisms by which polyQ modulate protein interactions might be the expansion of sequence-adjacent coiled-coil regions upon interaction of the coiled-coil region with another protein (Fiumara et al., 2010).

In support of this view, we found a strong association between polyQ and coiled-coil regions: both are found in the same sequence more often than random expectation, overlapping or at very short distance, as well as in proteins that interact with each other. This association was more significant than the association of polyQ to any protein domain. This includes the association to disordered regions (Simon and Hancock, 2009) that was weaker than the one found for coiled-coil regions (3.3-fold versus 6-fold enrichment in proximity of polyQ). In summary, our results indicate that polyQ expansions are selected in evolution to extend coiled-coil regions that take part in protein-protein interactions.

We found a strong bias for coiled-coil regions to be situated N-terminally of polyQ tracts. At the same time, polyP is sometimes found near polyQ and if so, often C-terminally to the polyQ tract. This is in agreement with the finding that polyP stabilizes

the structure of adjacent polyQ when located C-terminally but not when located N-terminally of it (Bhattacharyya et al., 2006). In an X-ray study of huntingtin exon 1, having polyP C-terminally adjacent to polyQ, the polyP was found to adopt a classical left-handed poly-proline helix structure (Kim et al., 2009). In synthetic polyQ peptides polyP fused to the C-terminus was shown to force polyQ into such a helical structure as well (Darnell et al., 2007). The property of proline to influence conformation is known in contexts other than polyQ proteins (Williams et al., 2004; Wood et al., 1995).

The second most frequent amino acid we observed enriched in proximity of polyQ was histidine. Several studies reported a protective role for histidine, similar to that of proline, preventing aggregation of polyQ stretches (Jayaraman et al., 2009; Sen et al., 2003). We find it interesting to note that the third most strongly enriched amino acid in and around polyQ, alanine, is often found in the hydrophobic interface of coiled-coils (Gromiha and Parry, 2004).

According to this evidence, we propose that polyQ tracts have a tendency to follow a coiled-coil region that they expand upon protein interaction. The conformational extension by polyQ of the coiled-coil region is then stabilized and paused by a capping sequence which, like polyP, acts directionally to stabilize and stop the growth of the helical region (Figure 5.9).

In summary, our results lead to the following general picture of the function of polyQ: its activity as a motif for protein interaction is tightly related to the length of the polyQ tract itself, the character of the sequences adjacent to it and to the concentration of interacting protein partners. We assume that the normal interplay of all these elements would lead to an enhanced, highly stable and specific interaction. However, the complexity of this system also suggests that small perturbations could lead to pathological interactions either with altered affinities or with different partners. The complex interaction of factors influencing the function of polyQ tracts perhaps explains why so many processes have been found to contribute to the pathomechanism of polyQ diseases including transcriptional dysregulation (Truant et al., 2007), RNA toxicity (Li et al., 2008), impairment of the ubiquitin-proteasome system (Bence et al., 2001; Chai et al., 1999; Waelter et al., 2001), mitochondrial dysfunction (Lin and Beal, 2006) and disturbed calcium signaling (Tang et al., 2005).

Our results also suggest that a given species may accumulate an abundance of polyQ proteins to modulate many protein interactions. However, this may come at no small expense: protein networks with abundant polyQ proteins may be in a delicate balance in which aggregation can occur depending on the concentration of many molecules. This balance might be lost in specific tissues and circumstances as mechanisms to keep protein

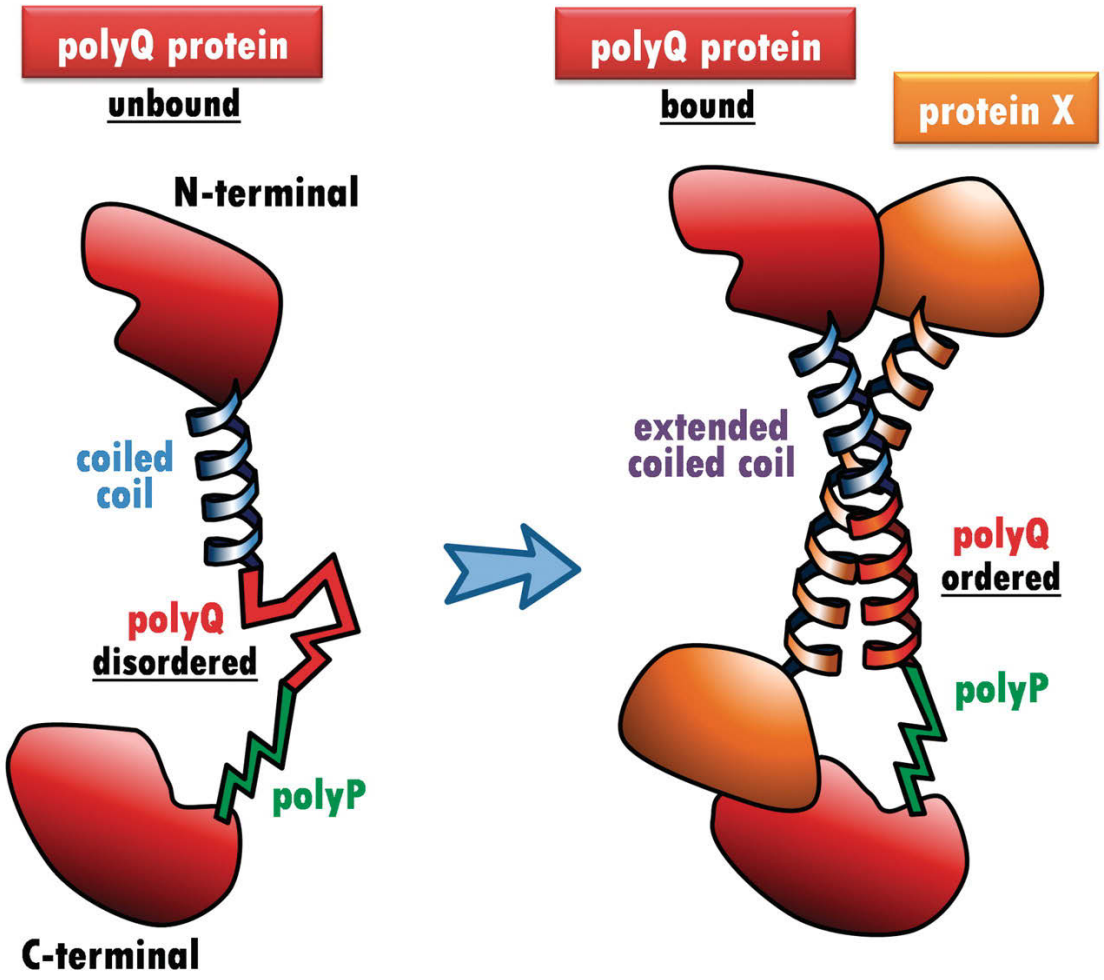

Figure 5.9: Cartoon of proposed polyQ function in protein interaction. Left: a polyQ protein contains a coiled-coil (blue), followed by a polyQ region (red) and a polyP region (green). In the unbound state, the polyQ region is disordered. Right: upon interaction with a protein partner X, the polyQ region adopts a coiled-coil structure that extends the original coiled-coil. The polyP region remains unstructured capping precisely the extension of the coiled-coil.

aggregates in check get challenged in ageing cells, as it has been observed in *C. elegans* (David et al., 2010). This may explain why neurons of the elderly are particular prone to anomalous polyQ expansion and in turn neurodegeneration.

We suggest that the study of the function of wild-type and pathogenic polyQ proteins will require experiments to test the variation in functionality that removing or expanding particular polyQ stretches will produce. Specifically, it needs to be investigated how these modifications influence the interaction abilities of the polyQ protein. Special attention should be paid to the gain or loss of interactions of other proteins with coiled-coil regions and polyQ. Explaining and predicting the effects of polyQ tracts will require elaborate analysis for each particular situation. The recent analysis of SCA1, where dramatic differences in the effects of the pathogenic protein were observed between brain regions (Jafar-Nejad et al., 2011), supports this idea.

In conclusion, our results show that the wild-type function of polyQ tracts is to modulate protein interactions depending on their molecular context. Therefore, the study of polyQ proteins will require correlating modification of this context to modifications in the protein interaction network.

5.7 Contributions

This chapter is a modified and extended version of Schaefer et al. (2012b). I did all the computational analyses under supervision of Miguel Andrade. The interpretation of the results was done together with Erich Wanker. The manuscript was written by Miguel Andrade and me. For this thesis, I largely extended the sections associating polyQ with PPIs after the publication of the manuscript.

6 Discussion

6.1 Selection of high-confidence and context-specific interactions

Proteins do not act alone but achieve diverse cellular functions in cooperation with other proteins. When these normal functions are interrupted, cells can end up in a disease state. Therefore, the knowledge of the complex maps of PPIs and tools for the interpretation of these data is fundamental for understanding of both normal cellular function and processes that lead to disease. Accordingly, many research efforts focus on the discovery of PPIs.

Accounting for the large number of diseases that is caused by perturbations of the cellular PPI network, new research disciplines develop disease network models and predict disease causing proteins from PPI data (Barabási et al., 2011). In network-based association studies genetic variation in a group of individuals carrying a disease is contrasted with knowledge on gene regulatory relations and physical interactions between proteins to elucidate pathways affected by a disease (Califano et al., 2012). Network pharmacology aims to identify novel drugs or drug targets based on analysis of PPI data (Hopkins, 2008). Therapeutic strategies are developed that inhibit disease-specific aberrant PPIs (Wells and McClendon, 2007; Zhao and Chmielewski, 2005). All of these approaches depend on the knowledge of reliable and representative PPI maps.

Reaching these goals is difficult given that many methods for detecting PPIs produce false positive measurements at large scales and fail to report many naturally occurring interactions. Additionally, many interactions that have been measured are strongly context-dependent and might occur only in a limited set of cell types or cellular components. Lastly, technical limitations and research interests introduce strong biases on the global interaction landscape reported so far.

In this thesis, we address these problems and present, at first, a method that assigns a continuous confidence score to reported PPIs that reflects the amount and quality of supporting experimental evidence. We believe that this is an essential contribution to help the network biology community to cope with the rapidly growing amount of PPI

data.

Though a few other approaches exist that integrate the experimental description for the purpose of scoring PPIs, our approach focuses on experimental parameters only. In comparison to the MINT score (Ceol et al., 2010), which is the closest to ours, we computationally optimized the selection of parameters of the scoring formula. Also, we developed a detailed quality assessment scheme that assigns an individual reliability value to each of more than 100 experimental methods that report PPIs, unlike MINT that weights experimental quality by how many interactions a method reports (high- versus low-throughput). Another difference is that MINT weights the number of studies by the number of times they have been cited. We believe that this further biases the score distribution towards well studied protein pairs rather than identifying reliable interactions. Together these differences might explain why our score outperforms the MINT score when ranking interactions in terms of reproducibility.

Other approaches to score PPIs integrate functional information of the protein pair. While we do see a correlation between the amount and quality of the experimental evidence and the specificity of the functional similarity (Chapter 4), we choose to separate these two concepts. Indeed, two proteins of the same pathway or transcriptions factors of the same class are more likely to interact but there are important naturally occurring interactions across functional classes. Accordingly, filtering a priori for functional similarity will bias the resulting network further towards homogeneous interactions between proteins that are functionally well characterized. Instead, we suggest to apply filters for experimental reliability and functional coherence independently, depending on the research question. We showed in Chapter 4 how the generation of tissue- or function-specific high-confidence networks can identify important causative interactions in disease.

An analysis that takes into account network topology (for example, the identification of hub proteins) or aims to estimate quality parameters of a novel screen using the overlap with existing PPI data (as suggested by Venkatesan et al. (2009)) depends on a maximally unbiased network. In contrast, in Chapter 3 we described a large bias resulting from non-uniform bait usage in the PPI network and observed interactions among well-studied proteins to have on average a higher confidence score. Accordingly, we see a limitation of our scoring approach in the fact that any selection of high-scoring interactions will further bias the resulting high-confidence network towards these frequently studied proteins (applying functional scoring schemes instead will have the same effect). In Chapter 5, we presented strategies to correct for the study bias and applied them to the analysis of polyQ-containing proteins to show that they have more interaction

partners than expected by chance. Indeed, correcting for the bias altered the outcome significantly. In summary, the set of interactions for a network analysis needs to be carefully chosen and requires one to choose the right balance between reliability and functional coherence on one side and degree of bias on the other.

Using a large integrated network we studied the effect on the network topology of merging thousands of studies. It is still under debate which distribution type would describe the degree distribution of PPI data best: the predominant view claims power-law distribution of the PPI degree but usually bases this decision only on visual inspection (Lima-Mendez and Van Helden, 2009). We find it interesting to note that for most experimental PPI networks a power-law distribution cannot be fitted sufficiently well (Khanin and Wit, 2006). For HIPPIE we observe a significant tendency (in the tail ≥ 45 interactions) to be power-law distributed. This could be partially caused by the almost perfectly power law distributed bait usage distribution in combination with the significant correlation between bait usage and degree of a protein. In the end, maybe the process of merging many networks introduces power law characteristics in integrated networks rather than evolutionary processes such as gene duplication, which is commonly believed to confer power-law properties to PPI networks (Bhan et al., 2002; Pastor-Satorras et al., 2003).

Besides the challenge of dealing with uncertainty and biases in PPI networks, the increasing amount of PPI data requires methods to identify interactions relevant to a particular biological problem. For example, not all PPIs of a disease-related protein contribute equally to disease development. Instead, a few alterations in the binding patterns of certain proteins provide the basis of the disease phenotype. For example, TP53 controls apoptosis upon DNA damage. Its activity is tightly controlled by different pathways and in cancer cells these control mechanisms are often interrupted. An important interaction that controls the activity of TP53 is with MDM2, which targets TP53 for degradation. Consequently, this interaction has been inhibited with small molecules (Klein and Vassilev, 2004). In Chapter 4, we presented a strategy to highlight interactions that might be relevant for disease mechanisms or in other non-disease contexts by incorporating functional annotation and gene expression information in combination with signal flow inferred from shortest paths connecting receptors with transcription factors. To our knowledge we are the first to combine network algorithms and context annotation to generate PPI networks of higher plausibility and to identify relevant interactions within them. We validate our approach by demonstrating that our method captures properties of canonical pathways, recovers known and novel disease mechanisms and leads to networks of higher experimental reliability.

Another way to use functional annotations in PPI networks is to correct for functional biases in the protein set being studied. We exemplified this in Chapter 5 for the detection of polyQ associated domains where we corrected for the bias towards transcription factors among polyQ proteins.

To make the integrated, context-associated and scored network data available, we implemented HIPPIE, a web tool that allows access to the constantly updated PPI network and provides numerous analysis options to conduct the graph mining tasks we have described here. Our approach of integrating PPI data with estimations of experimental reliability, functional and expression information provides several different views of the human PPI network:

- Disease networks can be studied under more realistic conditions by excluding PPIs that would not occur in the affected cell type.
- Research is guided by highlighting interactions between proteins associated to the research topic (e.g., between proteins involved in transcription or located in certain cellular compartments).
- Analysis of networks can be restricted to small numbers of highly reliable interactions.

The web tool HIPPIE combines many novel features that help experimentalists and likewise computational network biologists in the evaluation and analysis of PPI networks.

6.2 PolyQ function and disease

We used our functionally annotated network to provide important insights into the biological role of polyQ, a research question that has been discussed for decades: already more than 45 years ago structural features of glutamine-rich regions have been studied (Krull et al., 1965) and in the 1990s Nobel prize winner Max Perutz published a series of papers investigating the structure and function of polyQ (for example, (Perutz et al., 1993)). Our observations, which provide evidence that polyQ stabilizes PPIs and suggest that polyQ extends neighboring coiled-coil regions, are important contributions to the understanding of polyQ function. Even though the main focus of our investigation of polyQ function depends on network analyses, we observed that this cannot substantiate our hypothesis alone and must be combined with other analyses on the genomic and phylogenetic level. For example, observing more interaction partners indicates an involvement in PPIs, but only the association to PPI domains on sequence level produces

a consistent model explaining the mechanism by which polyQ stabilizes PPIs. Correlating polyQ presence at the genomic level with domains involved in aggregate clearance finally revealed the constraints by which polyQ evolution is driven: as a PPI domain its presence is beneficial but dangerous, since length expansion can drive the cell into disease state. Only when a protection machinery is present can this risk be taken.

Our model provides answers to several questions that have been puzzling researchers until now, like why transcription factors have more polyQ repeats than other proteins (because they are involved in more PPIs) and why prokaryotes have almost no proteins with polyQ repeats (because they do not have mechanisms to deal with aggregates formed by polyQ repeat proteins).

Ataxin-1, like Huntingtin, is a polyQ disease protein. In a recent collaboration project with Erich Wanker, we screened for modifiers of Ataxin-1 toxicity and observed a strong enrichment of coiled-coils in the set of proteins enhancing the aggregation of Ataxin-1 with length expanded polyQ (Petrakis et al., 2012). Removing the coiled-coil from one of these enhancing proteins, indeed reduced Ataxin-1 aggregation. In agreement with another recent study (Fiumara et al., 2010), these observations implicate coiled-coils in polyQ aggregation and thereby highlight the necessity of studying the wild type function of polyQ proteins for the understanding of the polyQ disease mechanisms.

6.3 Outlook

The significance of pre- and post-translational modification mechanisms for the modulation of the binding behaviour of a protein is becoming increasingly apparent: a recent study observed an enrichment of protein binding motifs in protein sequences encoded by alternatively spliced exons, which illustrates the importance of splicing for the modulation of PPIs (Weatheritt et al., 2012). Many protein interaction domains require post-translational modifications (PTMs) of the targeted binding motif. For example, the Src homology (SH2) domain binds phosphorylated tyrosine residues. Other PTMs that facilitate interactions are ubiquitination and acetylation (an overview of PTM-dependent PPIs is given in Seet et al. (2006)). Even though primary studies usually report on characteristics of the tested protein and methods are developed to specifically test for interactions conditional on PTMs (Wehr et al., 2008), PPI databases usually do not consider PTMs or disease mutations of the interacting protein pair. Only a few public databases report splice variants, mutations or PTMs in the proteins detected to interact (often containing only a small number of entries). This is partly due to a lack of standardization in the naming conventions for protein variants. As a start, UniProt

developed a numbering system for different protein isoforms. A unified naming convention that allows one to characterize any deviation from a canonical protein sequence would certainly help to build PPI resources that integrate information on protein variants and would, for example, allow studies on integrated datasets contrasting wild type with disease networks.

To detect highly specific interactions that are only realized under limited number of conditions and, hence, have never been measured, PPI experiments must be repeated under varying biological conditions. Again, there is a lack of database infrastructure to associate PPIs with these experimental conditions and so these experiments must be accompanied by the development of resources that indicate the context under which a PPI has been detected. We feel that the here presented tool HIPPIE is an important contribution towards this goal.

Appendix - Supplementary Tables

Table 1: Scores for experimental methods that detect PPIs.

Technique	PSI	Score
3 hybrid method	MI:0588	5
acetylation assay		7.5
Affinity Capture-Luminescence		5
Affinity Capture-MS		5
Affinity Capture-RNA		2
Affinity Capture-Western		5
affinity chromatography technology	MI:0004	5
affinity technology	MI:0400	5
anti baitcoimmunoprecipitation	MI:0006	5
anti tagcoimmunoprecipitation	MI:0007	5
antibody array	MI:0678	5
array technology	MI:0008	3
atomic force microscopy	MI:0872	9
beta galactosidase complementation	MI:0010	5
beta lactamase complementation	MI:0011	5
bimolecular fluorescence complementation	MI:0809	6
Biochemical	MI:0401	1
Biochemical Activity		5
bioluminescence resonance energy transfer	MI:0012	6
Biophysical	MI:0013	1
blue native page	MI:0276	3
chromatin immunoprecipitation assay	MI:0402	2
chromatography technology	MI:0091	1
circular dichroism	MI:0016	9
classical fluorescence spectroscopy	MI:0017	7.5
Co-crystal Structure		10
Co-fractionation		1
Co-localization		1
coimmunoprecipitation	MI:0019	5
colocalization by fluorescent probes cloning	MI:0021	1
colocalization by immunostaining	MI:0022	1
colocalization/visualisation technologies	MI:0023	1
comigration in gel electrophoresis	MI:0807	3

Continued on next page

Table 1 – continued

Technique	PSI	Score
comigration in non denaturing gel electrophoresis	MI:0404	3
comigration in sds page	MI:0808	3
competition binding	MI:0405	5
confocal microscopy	MI:0663	1
copurification	MI:0025	2
cosedimentation	MI:0027	2
cosedimentation in solution	MI:0028	2
cosedimentation through density gradient	MI:0029	2
cross-linking study	MI:0030	5
deacetylase assay	MI:0406	7.5
demethylase assay	MI:0870	7.5
dihydrofolatereductase reconstruction	MI:0111	6
dynamic light scattering	MI:0038	9
electron microscopy	MI:0040	5
electron paramagnetic resonance	MI:0042	9
electron tomography	MI:0410	9
electrophoretic mobility shift assay	MI:0413	2
electrophoretic mobility supershift assay	MI:0412	2
enzymatic study	MI:0415	1
enzyme linked immunosorbent assay	MI:0411	5
experimental interaction detection	MI:0045	1
far western blotting	MI:0047	5
filamentous phage display	MI:0048	6
filter binding	MI:0049	5
fluorescence correlation spectroscopy	MI:0052	10
fluorescence microscopy	MI:0416	1
fluorescence polarization spectroscopy	MI:0053	10
fluorescence technology	MI:0051	1
fluorescence-activated cell sorting	MI:0054	1
fluorescent resonance energy transfer	MI:0055	6
footprinting	MI:0417	3
FRET		6
gal4 vp16 complementation	MI:0728	5
genetic interference	MI:0254	0
gst pull down	MI:0059	5
gtpase assay	MI:0419	7.5
his pull down	MI:0061	5
homogeneous time resolved fluorescence	MI:0510	7
imaging technique	MI:0428	1
in vitro	MI:0492	1
in vivo	MI:0493	1

Continued on next page

Table 1 – continued

Technique	PSI	Score
in-gel kinase assay	MI:0423	7.5
inferred by curator	MI:0364	1
ion exchange chromatography	MI:0226	3
isothermal titration calorimetry	MI:0065	10
kinase homogeneous time resolved fluorescence	MI:0420	7.5
lambda phage display	MI:0066	6
lex-a dimerization assay	MI:0369	5
light microscopy	MI:0426	1
light scattering	MI:0067	10
mammalian protein protein interaction trap	MI:0231	6
mass spectrometry studies of complexes	MI:0069	5
methyltransferase assay	MI:0515	7.5
methyltransferase radiometric assay	MI:0516	7.5
molecular sieving	MI:0071	2
no experiment assigned		0
nuclear magnetic resonance	MI:0077	10
peptide array	MI:0081	5
phage display	MI:0084	6
phosphatase assay	MI:0434	7.5
phosphotransfer assay		7.5
polymerization	MI:0953	5
protease assay	MI:0435	7.5
protein array	MI:0089	5
protein complementation assay	MI:0090	5
protein cross-linking with a bifunctional reagent	MI:0031	5
protein kinase assay	MI:0424	7.5
protein tri hybrid	MI:0437	5
Protein-peptide		5
Protein-RNA		0
pull down	MI:0096	2.5
pull-down/mass spectrometry		5
Reconstituted Complex		10
reverse phase chromatography	MI:0227	1
reverse two hybrid	MI:0726	5
ribonuclease assay	MI:0920	7.5
saturation binding	MI:0440	7.5
scintillation proximity assay	MI:0099	7.5
solid phase assay	MI:0892	1
surface plasmon resonance	MI:0107	10
t7 phage display	MI:0108	6
tandem affinity purification	MI:0676	5

Continued on next page

Table 1 – continued

Technique	PSI	Score
tox-r dimerization assay	MI:0370	5
transcriptional complementation assay	MI:0232	5
transmission electron microscopy	MI:0020	5
two hybrid fragment pooling approach	MI:0399	5
Two-hybrid	MI:0018	5
ubiquitin reconstruction	MI:0112	5
x ray scattering	MI:0826	9
x-ray crystallography	MI:0114	10
x-ray fiber diffraction	MI:0825	9
yeast display	MI:0115	5

Table 2: Comprehensive network of influenza interference with cytokines. The directed networks were generated by computing shortest paths between viral proteins and genes upregulated upon influenza infection. Only viral proteins and host factors up to layer two are shown. Layer two proteins were required to be associated with cytokine-related pathways.

Viral proteins	First layer	Second layer	Tissues
PB1, PB2	BHLHE40	STAT3	Lung
PA	CDC42EP4	CDC42	Lung
PB2	CREB3	TAP1	Lung
NP	MAGED1	PJA1	Lung
NP	MAGED1	IRAK1	Lung
NP	MAGED1	HSPB1	Lung
NP	MAGED1	TOLLIP	Lung
PB1, PB2, M2	RBPMS	EWSR1	Lung
PB1, PB2, M2	RBPMS	TOLLIP	Lung
PA	RNF5	UBE2D2	Lung
PA	RNF5	UBE2D4	Lung
PA	RNF5	PXN	Lung
PA	RNF5	UBE2V1	Lung
PA	RNF5	UBE2Z	Lung
PA	RNF5	UBE2E3	Lung
PB1	SIAH1	STAT3	Lung
PB1	SIAH1	MYD88	Lung
PB1	SIAH1	UBE2E3	Lung
PB1	SIAH1	UBE2D2	Lung
PB1	TRIP6	SRC	Lung
PB1	TRIP6	NEDD9	Lung

Continued on next page

Table 2 – continued

Viral proteins	**First layer**	**Second layer**	**Tissues**
NS2	AIMP2	STUB1	Lung, BET
PB1, PB2	BHLHE40	TOLLIP	Lung, BET
PB2	CREB3	JUN	Lung, BET
PB2	CREB3	BCL2L1	Lung, BET
NS1, M1	PRKRA	SHC1	Lung, BET
NS1, M1	STAU1	CDC42	Lung, BET
NS1, M1	STAU1	RAC1	Lung, BET
PB1	TRIP6	PXN	Lung, BET
NS2	AIMP2	PPP2R1A	BET
PB1, PB2	BHLHE40	SUMO1	BET
PA	NDUFS3	MYC	BET
NS1, M1	PRKRA	RB1	BET
NS1, M1	STAU1	RPS6	BET
NS1, M1	STAU1	RPL6	BET
NS1, M1	STAU1	HDAC1	BET
NS1, M1	STAU1	UBE2D3	BET
NS1, M1	STAU1	FLNB	BET
PB1	TRIP6	PTK2	BET
PB1	TRIP6	FLNB	BET

Nomenclature

AD Activation domain

BD DNA-binding domain

GO Gene Ontology

HD Huntington's disease

MS Mass spectrometry

PI Phosphatidylinositol

PolyP Polyproline

PolyQ Polyglutamine

PPI Protein-protein interaction

PSI-MI Proteomics Standards Initiative Molecular Interaction

PSICQUIC Proteomics Standards Initiative Common Query Interface

PTM Post-translational modification

SILAC Stable isotope labeling by amino acids in cell culture

TAP Tandem affinity purification

UTR Untranslated regions

Y2H Yeast two-hybrid

Bibliography

Adachi, M., Matsukura, S., Tokunaga, H., and Kokubu, F. Expression of cytokines on human bronchial epithelial cells induced by influenza virus A. *Int Arch Allergy Appl Immunol*, 113(1-3):307–311, 1997.

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, 2006.

Agarwal, S., Deane, C. M., Porter, M. A., and Jones, N. S. Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS Comput Biol*, 6(6):e1000817, 2010.

Alba, M. M. and Guigo, R. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res*, 14(4):549–554, 2004.

Alba, M. M., Santibanez-Koref, M. F., and Hancock, J. M. The comparative genomics of polyglutamine repeats: extreme differences in the codon organization of repeat-encoding regions between mammals and Drosophila. *J Mol Evol*, 52(3):249–259, 2001.

Albers, M., Kranz, H., Kober, I., Kaiser, C., Klink, M., Suckow, J., Kern, R., and Koegl, M. Automated yeast two-hybrid screening for nuclear receptor-interacting proteins. *Mol Cell Proteomics*, 4(2):205–213, 2005.

Apweiler, R., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., Barrell, D., Bely, B., Bingley, M., Binns, D., Bower, L., Browne, P., Chan, W. M., Dimmer, E., Eberhardt, R., Fazzini, F., Fedotov, A., Foulger, R., Garavelli, J., Castro, L. G., Huntley, R., Jacobsen, J., Kleen, M., Laiho, K., Legge, D., Lin, Q., Liu, W., Luo, J., Orchard, S., Patient, S., Pichler, K., Poggioli, D., Pontikos, N., Pruess, M., Rosanoff, S., Sawford, T., Sehra, H., Turner, E., Corbett, M., Donnelly, M., van Rensburg, P., Xenarios, I., Bougueleret, L., Auchincloss, A., Argoud-Puy, G., Axelsen, K., Bairoch, A., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Bollondi, L., Boutet, E., Quintaje, S. B., Breuza, L., Bridge, A., DeCastro, E., Coudert, E., Cusin, I., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gehant, S., Ferro, S., Gasteiger, E., Gateau, A., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hulo, N., James, J., Jimenez, S., Jungo, F., Kappler, T., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Martin, X., Masson, P., Moinat, M., Morgat, A., Paesano, S., Pedruzzi, I., Pilbout, S., Poux, S., Pozzato, M., Redaschi, N., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stanley, E., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey,

A. L., Wu, C. H., Arighi, C. N., Arminski, L., Barker, W. C., Chen, C., Chen, Y., Dubey, P., Huang, H., Mazumder, R., McGarvey, P., Natale, D. A., Natarajan, T. G., Nchoutmboube, J., Roberts, N. V., Suzek, B. E., Ugochukwu, U., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S., and Zhang, J. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*, 39:D214–D219, 2011.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, 38(Database issue):D525–31, 2010.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S., Ceol, A., Chautard, E., Dana, J. M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R. E., Isserlin, R., Jimenez, R. C., Kerssemakers, J., Khadake, J., Lynn, D. J., Michaut, M., O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G. D., Cesareni, G., Donaldson, I. M., Eisenberg, D., Kleywegt, G. J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, M., and Hermjakob, H. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods*, 8(7):528–529, 2011.

Bader, G. and Hogue, C. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, 2003.

Bader, G. D., Betel, D., and Hogue, C. W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31(1):248–250, 2003.

Bandyopadhyay, S., Chiang, C. Y., Srivastava, J., Gersten, M., White, S., Bell, R., Kurschner, C., Martin, C. H., Smoot, M., Sahasrabudhe, S., Barber, D. L., Chanda, S. K., and Ideker, T. A human MAP kinase interactome. *Nat Methods*, 7(10):801–805, 2010.

Barabási, A. L. and Albert, R. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

Barabási, A. L. and Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, 2004.

Barabási, A. L., Gulbahce, N., and Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68, 2011.

Barbato, C., Corbi, N., Canu, N., Fanciulli, M., Serafino, A., Ciotti, M., Libri, V., Bruno, T., Amadoro, G., De Angelis, R., Calissano, P., and Passananti, C. Rb binding protein Che-1 interacts with Tau in cerebellar granule neurons: Modulation during neuronal apoptosis. *Mol Cell Neurosci*, 24(4):1038–1050, 2003.

Barbosa-Silva, A., Fontaine, J. F., Donnard, E. R., Stussi, F., Ortega, J. M., and Andrade-Navarro, M. A. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. *BMC Bioinformatics*, 12(1):435, 2011.

Barrios-Rodiles, M., Brown, K. R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I. W., Luga, V., Przulj, N., Robinson, M., Suzuki, H., Hayashizaki, Y., Jurisica, I., and Wrana, J. L. High-throughput mapping of a dynamic signaling network in mammalian cells. *Sci STKE*, 307(5715):1621, 2005.

Behrends, C., Sowa, M. E., Gygi, S. P., and Harper, J. W. Network organization of the human autophagy system. *Nature*, 466(7302):68–76, 2010.

Belisle, S. E., Tisoncik, J. R., Korth, M. J., Carter, V. S., Proll, S. C., Swayne, D. E., Pantin-Jackwood, M., Tumpey, T. M., and Katze, M. G. Genomic Profiling of Tumor Necrosis Factor Alpha (TNF-α) Receptor and Interleukin-1 Receptor Knockout Mice Reveals a Link between TNF-α Signaling and Increased Severity of 1918 Pandemic Influenza Virus Infection. *J Virol*, 84(24):12576–12588, 2010.

Bell, R., Hubbard, A., Chettier, R., Chen, D., Miller, J. P., Kapahi, P., Tarnopolsky, M., Sahasrabuhde, S., Melov, S., and Hughes, R. E. A human protein interaction network shows conservation of aging processes between human and invertebrate species. *PLoS Genet*, 5(3):e1000414, 2009.

Bence, N. F., Sampat, R. M., and Kopito, R. R. Impairment of the ubiquitin-proteasome system by protein aggregation. *Science*, 292(5521):1552–1555, 2001.

Bhan, A., Galas, D. J., and Dewey, T. G. A duplication growth model of gene expression networks. *Bioinformatics*, 18(11):1486–1493, 2002.

Bhattacharyya, A., Thakur, A. K., Chellgren, V. M., Thiagarajan, G., Williams, A. D., Chellgren, B. W., Creamer, T. P., and Wetzel, R. Oligoproline effects on polyglutamine conformation and aggregation. *J Mol Biol*, 355(3):524–535, 2006.

Björklund, A. K., Light, S., Hedin, L., and Elofsson, A. Quantitative assessment of the structural bias in protein–protein interaction assays. *Proteomics*, 8(22):4657–4667, 2008.

Bossi, A. and Lehner, B. Tissue specificity and the human protein interaction network. *Mol Syst Biol*, 5:260, 2009.

Boutell, J. M., Thomas, P., Neal, J. W., Weston, V. J., Duce, J., Harper, P. S., and Jones, A. L. Aberrant interactions of transcriptional repressor proteins with the Huntington's disease gene product, huntingtin. *Hum Mol Genet*, 8(9):1647–1655, 1999.

Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P. O., Bergamini, G., Croughton, K., Cruciat, C., Eberhard, D., Gagneur, J., Ghidelli, S., Hopf, C., Huhse, B., Mangano, R., Michon, A. M., Schirle, M., Schlegl, J., Schwab, M., Stein, M. A., Bauer, A., Casari, G., Drewes, G., Gavin, A. C., Jackson, D. B., Joberty, G., Neubauer, G., Rick,

J., Kuster, B., and Superti-Furga, G. A physical and functional map of the human TNF-α/NF-κB signal transduction pathway. *Nat Cell Biol*, 6(2):97–105, 2004.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Roncari, L., de Smet, A. S., Venkatesan, K., Rual, J. F., Vandenhaute, J., Cusick, M. E., Pawson, T., Hill, D. E., Tavernier, J., Wrana, J. L., Roth, F. P., and Vidal, M. An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*, 6(1):91–97, 2008.

Brown, K. R. and Jurisica, I. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, 2005.

Bruno, T., Desantis, A., Bossi, G., Di Agostino, S., Sorino, C., De Nicola, F., Iezzi, S., Franchitto, A., Benassi, B., Galanti, S., La Rosa, F., Floridi, A., Bellacosa, A., Passananti, C., Blandino, G., and Fanciulli, M. Che-1 promotes tumor cell survival by sustaining mutant p53 transcription and inhibiting DNA damage response activation. *Cancer Cell*, 18(2):122–134, 2010.

Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, 433(7025):531–537, 2005.

Butland, S. L., Devon, R. S., Huang, Y., Mead, C. L., Meynert, A. M., Neal, S. J., Lee, S. S., Wilkinson, A., Yang, G. S., Yuen, M. M., Hayden, M. R., Holt, R. A., Leavitt, B. R., and Ouellette, B. F. CAG-encoded polyglutamine length polymorphism in the human genome. *BMC Genomics*, 8:126, 2007.

Califano, A., Butte, A. J., Friend, S., Ideker, T., and Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat Genet*, 44(8):841–847, 2012.

Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., and Lowry, S. F. A network-based analysis of systemic inflammation in humans. *Nature*, 437(7061): 1032–1037, 2005.

Cario, E. and Podolsky, D. K. Intestinal epithelial TOLLerance versus inTOLLerance of commensals. *Mol Immunol*, 42(8):887–893, 2005.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res*, 38(Database issue):D532–9, 2010.

Chai, Y., Koppenhafer, S. L., Shoesmith, S. J., Perez, M. K., and Paulson, H. L. Evidence for proteasome involvement in polyglutamine disease: localization to nuclear inclusions

in SCA3/MJD and suppression of polyglutamine aggregation in vitro. *Hum Mol Genet*, 8(4):673–682, 1999.

Chai, Y., Wu, L., Griffin, J. D., and Paulson, H. L. The role of protein composition in specifying nuclear inclusion formation in polyglutamine disease. *J Biol Chem*, 276 (48):44889–44897, 2001.

Chai, Y., Shao, J., Miller, V. M., Williams, A., and Paulson, H. L. Live-cell imaging reveals divergent intracellular dynamics of polyglutamine disease proteins and supports a sequestration model of pathogenesis. *Proc Natl Acad Sci U S A*, 99(14):9310–9315, 2002.

Chan, P. P. and Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res*, 37(Database issue):D93–7, 2009.

Chaurasia, G., Iqbal, Y., Hanig, C., Herzel, H., Wanker, E. E., and Futschik, M. E. UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res*, 35 (Database issue):D590–4, 2007.

Chen, J. Y., Shen, C., and Sivachenko, A. Y. Mining Alzheimer disease relevant proteins from integrated protein interactome data. In *Pac Symp Biocomput*, volume 11, pages 367–378, 2006.

Chiti, F. and Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem*, 75:333–366, 2006.

Chowdhary, R., Tan, S. L., Zhang, J., Karnik, S., Bajic, V. B., and Liu, J. S. Context-Specific Protein Network Miner–An Online System for Exploring Context-Specific Protein Interaction Networks from the Literature. *PloS One*, 7(4):e34480, 2012.

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140, 2007.

Chun, W. and Johnson, G. V. The role of tau phosphorylation and cleavage in neuronal cell death. *Front Biosci*, 12:733–756, 2007.

Claudio, P. and Maurizio, F. The anti-apoptotic factor Che-1/AATF links transcriptional regulation, cell cycle control, and DNA damage response. *Cell Division*, 2(1): 21, 2007.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Colland, F., Jacq, X., Trouplin, V., Mougin, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P., and Gauthier, J. M. Functional proteomics mapping of a human signaling pathway. *Genome Res*, 14(7):1324–1332, 2004.

Cummings, C. J., Mancini, M. A., Antalffy, B., DeFranco, D. B., Orr, H. T., and Zoghbi, H. Y. Chaperone suppression of aggregation and altered subcellular proteasome localization imply protein misfolding in SCA1. *Nat Genet*, 19(2):148–154, 1998.

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A. R., Simonis, N., Rual, J. F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P., and Vidal, M. Literature-curated protein interaction datasets. *Nat Methods*, 6(1):39–46, 2008.

Darnell, G., Orgel, J., Pahl, R., and Meredith, S. C. Flanking Polyproline Sequences Inhibit [beta]-Sheet Structure in Polyglutamine Segments by Inducing PPII-like Helix Structure. *J Mol Biol*, 374(3):688–704, 2007.

David, D. C., Ollikainen, N., Trinidad, J. C., Cary, M. P., Burlingame, A. L., and Kenyon, C. Widespread protein aggregation as an inherent part of aging in C. elegans. *PLoS Biol*, 8(8):e1000450, 2010.

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. Dynamic complex formation during the yeast cell cycle. *Sci STKE*, 307(5710):724, 2005.

Dehay, B. and Bertolotti, A. Critical role of the proline-rich region in Huntingtin for aggregation and cytotoxicity in yeast. *J Biol Chem*, 281(47):35608–35615, 2006.

Deng, M., Mehta, S., Sun, F., and Chen, T. Inferring domain–domain interactions from protein–protein interactions. *Genome Res*, 12(10):1540–1548, 2002.

Dennis, G. J., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4(5):P3, 2003.

Dickerson, J., Pinney, J., and Robertson, D. The biological context of HIV-1 host interactions reveals subtle insights into a system hijack. *BMC Systems Biology*, 4(1): 80, 2010.

Diebold, S. S., Kaisho, T., Hemmi, H., Akira, S., and e Sousa, C. Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. *Sci STKE*, 303(5663):1529, 2004.

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R., Gardner, M., Laiho, K., Legge, D., Magrane, M., Pichler, K., Poggioli, D., Sehra, H., Auchincloss, A., Axelsen, K., Blatter, M. C., Boutet, E., Braconi-Quintaje, S., Breuza, L., Bridge, A., Coudert, E., Estreicher, A., Famiglietti, L., Ferro-Rojas, S., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., James, J., Jimenez, S., Jungo, F., Keller, G., Lemercier, P., Lieberherr, D., Masson, P., Moinat, M., Pedruzzi, I., Poux, S., Rivoire, C., Roechert, B., Schneider, M., Stutz, A., Sundaram, S., Tognolli, M., Bougueleret, L., Argoud-Puy, G., Cusin, I., Duek-Roggli, P., Xenarios, I., and Apweiler, R. The

UniProt-GO Annotation database in 2011. *Nucleic Acids Res*, 40(D1):D565–D570, 2012.

DiNitto, J. P. and Lambright, D. G. Membrane and juxtamembrane targeting by PH and PTB domains. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1761(8):850–867, 2006.

Dolan, P. J. and Johnson, G. V. W. The role of tau kinases in Alzheimer's disease. *Curr Opin Drug Discov Devel*, 13(5):595, 2010.

Ehrhardt, C., Seyer, R., Hrincius, E. R., Eierhoff, T., Wolff, T., and Ludwig, S. Interplay between influenza A virus and the innate immune signaling. *Microbes Infect*, 12(1): 81–87, 2010.

Eichinger, L., Pachebat, J. A., Glockner, G., Rajandream, M. A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., Tunggal, B., Kummerfeld, S., Madera, M., Konfortov, B. A., Rivero, F., Bankier, A. T., Lehmann, R., Hamlin, N., Davies, R., Gaudet, P., Fey, P., Pilcher, K., Chen, G., Saunders, D., Sodergren, E., Davis, P., Kerhornou, A., Nie, X., Hall, N., Anjard, C., Hemphill, L., Bason, N., Farbrother, P., Desany, B., Just, E., Morio, T., Rost, R., Churcher, C., Cooper, J., Haydock, S., van Driessche, N., Cronin, A., Goodhead, I., Muzny, D., Mourier, T., Pain, A., Lu, M., Harper, D., Lindsay, R., Hauser, H., James, K., Quiles, M., Madan Babu, M., Saito, T., Buchrieser, C., Wardroper, A., Felder, M., Thangavelu, M., Johnson, D., Knights, A., Loulseged, H., Mungall, K., Oliver, K., Price, C., Quail, M. A., Urushihara, H., Hernandez, J., Rabbinowitsch, E., Steffen, D., Sanders, M., Ma, J., Kohara, Y., Sharp, S., Simmonds, M., Spiegler, S., Tivey, A., Sugano, S., White, B., Walker, D., Woodward, J., Winckler, T., Tanaka, Y., Shaulsky, G., Schleicher, M., Weinstock, G., Rosenthal, A., Cox, E. C., Chisholm, R. L., Gibbs, R., Loomis, W. F., Platzer, M., Kay, R. R., Williams, J., Dear, P. H., Noegel, A. A., Barrell, B., and Kuspa, A. The genome of the social amoeba Dictyostelium discoideum. *Nature*, 435 (7038):43–57, 2005.

Eisenberg, E. and Levanon, E. Y. Human housekeeping genes are compact. *Trends Genet*, 19(7):362–365, 2003.

Elkon, R., Vesterman, R., Amit, N., Ulitsky, I., Zohar, I., Weisz, M., Mass, G., Orlev, N., Sternberg, G., Blekhman, R., Assa, J., Shiloh, Y., and Shamir, R. SPIKE–a database, visualization and analysis tool of cellular signaling pathways. *BMC Bioinformatics*, 9 (1):110, 2008.

Estojak, J., Brent, R., and Golemis, E. A. Correlation of two-hybrid affinity data with in vitro measurements. *Mol Cell Biol*, 15(10):5820–5829, 1995.

Faux, N. G., Bottomley, S. P., Lesk, A. M., Irving, J. A., Morrison, J. R., de la Banda, M. G., and Whisstock, J. C. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res*, 15(4):537–551, 2005.

Fields, B. N., Knipe, D. M., and Howley, P. M. Fields' virology (5th), 2007.

Fields, S. and Song, O. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, July 1989.

Filimonenko, M., Isakson, P., Finley, K. D., Anderson, M., Jeong, H., Melia, T. J., Bartlett, B. J., Myers, K. M., Birkeland, H. C., Lamark, T., Krainc, D., Brech, A., Stenmark, H., Simonsen, A., and Yamamoto, A. The selective macroautophagic degradation of aggregated proteins requires the PI3P-binding protein Alfy. *Mol Cell*, 38(2):265–279, 2010.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A. The Pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–22, 2010.

Fiumara, F., Fioriti, L., Kandel, E. R., and Hendrickson, W. A. Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell*, 143 (7):1121–1135, 2010.

Fontaine, J. F., Barbosa-Silva, A., Schaefer, M., Huska, M. R., Muro, E. M., and Andrade-Navarro, M. A. MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res*, 37(suppl 2):W141–W146, 2009.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., Jacq, B., Arpin, M., Bellaiche, Y., Bellusci, S., Benaroch, P., Bornens, M., Chanet, R., Chavrier, P., Delattre, O., Doye, V., Fehon, R., Faye, G., Galli, T., Girault, J. A., Goud, B., de Gunzburg, J., Johannes, L., Junier, M. P., Mirouse, V., Mukherjee, A., Papadopoulo, D., Perez, F., Plessis, A., Rosse, C., Saule, S., Stoppa-Lyonnet, D., Vincent, A., White, M., Legrain, P., Wojcik, J., Camonis, J., and Daviet, L. Protein interaction mapping: a Drosophila case study. *Genome Res*, 15(3):376–384, 2005.

Friedman, M. J., Shah, A. G., Fang, Z. H., Ward, E. G., Warren, S. T., Li, S., and Li, X. J. Polyglutamine domain modulates the TBP-TFIIB interaction: implications for its normal function and neurodegeneration. *Nat Neurosci*, 10(12):1519–1528, 2007.

Futschik, M. E., Chaurasia, G., and Herzel, H. Comparison of human protein–protein interaction maps. *Bioinformatics*, 23(5):605–611, 2007.

Gandhi, T. K., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–293, 2006.

Gatchel, J. R. and Zoghbi, H. Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet*, 6(10):743–755, 2005.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084): 631–636, 2006.

Geeraedts, F., Goutagny, N., Hornung, V., Severa, M., De Haan, A., Pool, J., Wilschut, J., Fitzgerald, K. A., and Huckriede, A. Superior immunogenicity of inactivated whole virus H5N1 influenza vaccine is primarily controlled by Toll-like receptor signalling. *PLoS Pathog*, 4(8):e1000138, 2008.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N. J., Millar, A. H., and Geisler, M. A predicted interactome for Arabidopsis. *Plant Physiol*, 145(2):317–329, 2007.

Giorgini, F. and Muchowski, P. J. Connecting the dots in Huntington's disease with protein interaction networks. *Genome Biol*, 6(3):210, 2005.

Gissi, C., Pesole, G., Cattaneo, E., and Tartari, M. Huntingtin gene evolution in Chordata and its peculiar features in the ascidian Ciona genus. *BMC Genomics*, 7:288, 2006.

Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., Droege, A., Lindenberg, K. S., Knoblich, M., Haenig, C., Herbst, M., Suopanki, J., Scherzinger, E., Abraham, C., Bauer, B., Hasenbank, R., Fritzsche, A., Ludewig, A. H., Bussow, K., Coleman, S. H., Gutekunst, C. A., Landwehrmeyer, B. G., Lehrach, H., and Wanker, E. E. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell*, 15(6):853–865, 2004.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A. L. The human disease network. *Proc Natl Acad Sci U S A*, 104(21):8685, 2007.

Gottipati, S., Rao, N. L., and Fung-Leung, W. P. IRAK1: a critical signaling mediator of innate immunity. *Cell Signal*, 20(2):269–276, 2008.

Gromiha, M. M. and Parry, D. A. D. Characteristic features of amino acid residues in coiled-coil protein structures. *Biophys Chem*, 111(2):95–103, 2004.

Guimarães, K. S., Jothi, R., Zotenko, E., and Przytycka, T. M. Predicting domain-domain interactions using a parsimony approach. *Genome Biol*, 7(11):R104, 2006.

Hagberg, A., Swart, P., and S Chult, D. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Laboratory (LANL), 2008.

Hale, B. G., Randall, R. E., Ortín, J., and Jackson, D. The multifunctional NS1 protein of influenza A viruses. *J Gen Virol*, 89(10):2359–2376, 2008.

Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P., and Vidal, M. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430 (6995):88–93, 2004.

Hancock, J. M. The contribution of slippage-like processes to genome evolution. *J Mol Evol*, 41(6):1038–1047, 1995.

Hands, S., Sinadinos, C., and Wyttenbach, A. Polyglutamine gene function and dysfunction in the ageing brain. *Biochim Biophys Acta*, 1779(8):507–521, 2008.

Harrison, P. M. Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila. *BMC Bioinformatics*, 7:441, 2006.

Hart, G. T., Ramani, A. K., and Marcotte, E. M. How complete are current yeast and human protein-interaction networks. *Genome Biol*, 7(11):120, 2006.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. From molecular to modular cell biology. *Nature*, 402(6761):47, 1999.

He, Y., Xu, K., Keiner, B., Zhou, J., Czudai, V., Li, T., Chen, Z., Liu, J., Klenk, H. D., Shu, Y. L., and Sun, B. Influenza A virus replication induces cell cycle arrest in G0/G1 phase. *J Virol*, 84(24):12832–12840, 2010.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–183, 2004.

Hoey, T., Weinzierl, R. O., Gill, G., Chen, J. L., Dynlacht, B. D., and Tjian, R. Molecular cloning and functional analysis of Drosophila TAF110 reveal properties expected of coactivators. *Cell*, 72(2):247–260, 1993.

Holmes, S. E., Hearn, E. O., Ross, C. A., and Margolis, R. L. SCA12: an unusual mutation leads to an unusual spinocerebellar ataxia. *Brain Res Bull*, 56(3-4):397–403, 2001.

Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*, 4(11):682–690, 2008.

Huntley, M. and Golding, G. B. Evolution of simple sequence in proteins. *J Mol Evol*, 51(2):131–140, 2000.

Ignatova, Z. and Gierasch, L. M. Extended polyglutamine tracts cause aggregation and structural perturbation of an adjacent beta barrel protein. *J Biol Chem*, 281(18): 12959–12967, 2006.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569, 2001.

Ivanic, J., Yu, X., Wallqvist, A., and Reifman, J. Influence of protein abundance on high-throughput protein-protein interaction detection. *PloS One*, 4(6):e5815, 2009.

Ivanov, S. V., Salnikow, K., Ivanova, A. V., Bai, L., and Lerman, M. I. Hypoxic repression of STAT1 and its downstream genes by a pVHL/HIF-1 target DEC1/STRA13. *Oncogene*, 26(6):802–812, 2006.

Jafar-Nejad, P., Ward, C. S., Richman, R., Orr, H. T., and Zoghbi, H. Y. Regional rescue of spinocerebellar ataxia type 1 phenotypes by 14-3-3\varepsilon haploinsufficiency in mice underscores complex pathogenicity in neurodegeneration. *Proc Natl Acad Sci U S A*, 108(5):2142–2147, 2011.

Jayaraman, M., Kodali, R., and Wetzel, R. The impact of ataxin-1-like histidine insertions on polyglutamine aggregation. *Protein Eng Des Sel*, 22(8):469–478, 2009.

Jensen, L. J. and Bork, P. Not Comparable, But Complementary. *Science*, 322(5898): 56–57, 2008.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue):D412–6, 2009.

Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

Jeronimo, C., Forget, D., Bouchard, A., Li, Q., Chua, G., Poitras, C., Therien, C., Bergeron, D., Bourassa, S., Greenblatt, J., Chabot, B., Poirier, G. G., Hughes, T. R., Blanchette, M., Price, D. H., and Coulombe, B. Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. *Mol Cell*, 27(2):262–274, 2007.

Jonsson, P. F. and Bates, P. A. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297, 2006.

Kaltenbach, L. S., Romero, E., Becklin, R. R., Chettier, R., Bell, R., Phansalkar, A., Strand, A., Torcassi, C., Savage, J., Hurlburt, A., Cha, G. H., Ukani, L., Chepanoske, C. L., Zhen, Y., Sahasrabudhe, S., Olson, J., Kurschner, C., Ellerby, L. M., Peltier, J. M., Botas, J., and Hughes, R. E. Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genet*, 3(5):e82, 2007.

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res*, 39(suppl 1):D712–D717, 2011.

Karlin, S. and Burge, C. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc Natl Acad Sci U S A*, 93(4):1560–1565, 1996.

Karlin, S., Brocchieri, L., Bergman, A., Mrázek, J., and Gentles, A. J. Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci U S A*, 99(1): 333, 2002.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., and Kent, W. J. The UCSC Genome Browser Database. *Nucleic Acids Res*, 31(1):51–54, 2003.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, 40(D1):D841—-D846, 2012.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. Human Protein Reference Database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–72, 2009.

Khabar, K. S. A. The AU-rich transcriptome: more than interferons and cytokines, and its role in disease. *J Interferon Cytokine Res*, 25(1):1–10, 2005.

Khanin, R. and Wit, E. How scale-free are biological networks. *J Comput Biol*, 13(3): 810–818, 2006.

Kim, M. W., Chelliah, Y., Kim, S. W., Otwinowski, Z., and Bezprozvanny, I. Secondary structure of Huntingtin amino-terminal region. *Structure*, 17(9):1205–1212, 2009.

Klein, C. and Vassilev, L. T. Targeting the p53–MDM2 interaction to treat cancer. *Br J Cancer*, 91(8):1415–1419, 2004.

Kopito, R. R. Aggresomes, inclusion bodies and protein aggregation. *Trends Cell Biol*, 10(12):524–530, 2000.

Kovtun, I. V. and McMurray, C. T. Features of trinucleotide repeat instability in vivo. *Cell Res*, 18(1):198–213, 2008.

Kozlowski, P., de Mezer, M., and Krzyzosiak, W. J. Trinucleotide repeats in human genome and exome. *Nucleic Acids Res*, 38(12):4027–4039, 2010.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrin-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

Krull, L. H., Wall, J. S., Zobel, H., and Dimler, R. J. Synthetic Polypeptides Containing Side-Chain Amide Groups: Water-insoluble Polymers*. *Biochemistry (Mosc)*, 4(4): 626–633, 1965.

Kuemmerle, S., Gutekunst, C. A., Klein, A. M., Li, X. J., Li, S. H., Beal, M. F., Hersch, S. M., and Ferrante, R. J. Huntington aggregates may not predict neuronal death in Huntington's disease. *Ann Neurol*, 46(6):842–849, 1999.

LaPointe, N. E., Morfini, G., Pigino, G., Gaisina, I. N., Kozikowski, A. P., Binder, L. I., and Brady, S. T. The amino terminus of tau inhibits kinesin-dependent axonal transport: Implications for filament toxicity. *J Neurosci Res*, 87(2):440–451, 2009.

Lee, S. A., Chan, C. H., Chen, T. C., Yang, C. Y., Huang, K. C., Tsai, C. H., Lai, J. M., Wang, F. S., Kao, C. Y., and Huang, C. Y. F. POINeT: protein interactome with sub-network analysis and hub prioritization. *BMC Bioinformatics*, 10(1):114, 2009.

Lehner, B. and Fraser, A. G. A first-draft human protein-interaction map. *Genome Biol*, 5(9):R63, 2004.

Lehner, B. and Sanderson, C. M. A protein interaction framework for human mRNA degradation. *Genome Res*, 14(7):1315–1323, 2004.

Letunic, I., Doerks, T., and Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res*, 37(Database issue):D229–32, 2009.

Li, C., Bankhead III, A., Eisfeld, A. J., Hatta, Y., Jeng, S., Chang, J. H., Aicher, L. D., Proll, S., Ellis, A. L., Law, G. L., Li, C., Bankhead, A., Eisfeld, A. J., Hatta, Y., Jeng, S., Chang, J. H., Aicher, L. D., Proll, S., Ellis, A. L., Law, G. L., Waters, K. M., Neumann, G., Katze, M. G., McWeeney, S., and Kawaoka, Y. Host regulatory network response to infection with highly pathogenic h5n1 avian influenza virus. *J Virol*, 85(21):10955–10967, 2011.

Li, L. B., Yu, Z., Teng, X., and Bonini, N. M. RNA toxicity is a component of ataxin-3 degeneration in Drosophila. *Nature*, 453(7198):1107–1111, 2008.

Li, P., Huey-Tubman, K. E., Gao, T., Li, X., West, A. P., Bennett, M. J., and Bjorkman, P. J. The structure of a polyQ–anti-polyQ complex reveals binding according to a linear lattice model. *Nat Struct Mol Biol*, 14(5):381–387, 2007.

Li, S. H. and Li, X. J. Huntingtin–protein interactions and the pathogenesis of Huntington's disease. *Trends Genet*, 20(3):146–154, 2004.

Liang, Q., Luo, J., Zhou, K., Dong, J., and He, H. Immune-related gene expression in response to H5N1 avian influenza virus infection in chicken and duck embryonic fibroblasts. *Mol Immunol*, 48(6):924–930, 2011.

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabo, G., Rual, J. F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., Barabasi, A. L., Vidal, M., and Zoghbi, H. Y. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, 125(4):801–814, 2006.

Lima-Mendez, G. and Van Helden, J. The powerful law of the power law and other myths in network biology. *Mol Biosyst*, 5(12):1482–1493, 2009.

Lin, M. T. and Beal, M. F. Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature*, 443(7113):787–795, 2006.

Lin, Y., Dent, S. Y., Wilson, J. H., Wells, R. D., and Napierala, M. R loops stimulate genetic instability of CTG.CAG repeats. *Proc Natl Acad Sci U S A*, 107(2):692–697, 2010.

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., and Bader, G. D. Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26(18): 2347–2348, 2010.

Lopes, T. J. S., Schaefer, M., Shoemaker, J., Matsuoka, Y., Neumann, G., Andrade-Navarro, M. A., Kawaoka, Y., and Kitano, H. Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics*, 27(17):2414–2421, 2011.

Lu, L. J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*, 15(7):945–953, 2005.

Ludwig, S., Pleschka, S., Planz, O., and Wolff, T. Ringing the alarm bells: signalling and apoptosis in influenza virus infected cells. *Cell Microbiol*, 8(3):375–386, 2006.

Lund, J. M., Alexopoulou, L., Sato, A., Karow, M., Adams, N. C., Gale, N. W., Iwasaki, A., and Flavell, R. A. Recognition of single-stranded RNA viruses by Toll-like receptor 7. *Proc Natl Acad Sci U S A*, 101(15):5598, 2004.

Lupas, A., Van Dyke, M., and Stock, J. Predicting coiled coils from protein sequences. *Science*, 252(5009):1162–1164, 1991.

Maetschke, S. R., Simonsen, M., Davis, M. J., and Ragan, M. A. Gene Ontology-driven inference of protein–protein interactions using inducers. *Bioinformatics*, 28(1):69–75, 2012.

Magrane, M. and UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, 2011:bar009, 2011.

Mah, N., Wang, Y., Liao, M. C., Prigione, A., Jozefczuk, J., Lichtner, B., Wolfrum, K., Haltmeier, M., Flottmann, M., Schaefer, M., Hahn, A., Mrowka, R., Klipp, E., Andrade-Navarro, M. A., and Adjaye, J. Molecular Insights into Reprogramming-Initiation Events Mediated by the OSKM Gene Regulatory Network. *PloS One*, 6(8): e24351, 2011.

Major, M. B., Camp, N. D., Berndt, J. D., Yi, X., Goldenberg, S. J., Hubbert, C., Biechele, T. L., Gingras, A. C., Zheng, N., Maccoss, M. J., Angers, S., and Moon, R. T. Wilms tumor suppressor wtx negatively regulates wnt/{beta}-catenin signaling. *Sci STKE*, 316(5827):1043, 2007.

Mann, M. Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol*, 7(12):952–958, 2006.

Martin, A., Ochagavia, M. E., Rabasa, L. C., Miranda, J., Fernandez-de Cossio, J., and Bringas, R. BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics*, 11(1):91, 2010.

Matsukura, S., Kokubu, F., Noda, H., Tokunaga, H., and Adachi, M. Expression of IL-6, IL-8, and RANTES on human bronchial epithelial cells, NCI-H292, induced by influenza virus A. *J Allergy Clin Immunol*, 98(6):1080–1087, 1996.

McDonnell, A. V., Jiang, T., Keating, A. E., and Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*, 22(3):356–358, 2006.

Mi, K., Dolan, P. J., and Johnson, G. V. W. The low density lipoprotein receptor-related protein 6 interacts with glycogen synthase kinase 3 and attenuates activity. *J Biol Chem*, 281(8):4787–4794, 2006.

Miettinen, M., Sareneva, T., Julkunen, I., and Matikainen, S. IFNs activate toll-like receptor gene expression in viral infections. *Genes Immun*, 2(6):349, 2001.

Miller, B. W., Lau, G., Grouios, C., Mollica, E., Barrios-Rodiles, M., Liu, Y., Datti, A., Morris, Q., Wrana, J. L., and Attisano, L. Application of an integrated physical and functional screening approach to identify inhibitors of the Wnt pathway. *Mol Syst Biol*, 5:315, 2009.

Miller, J. W., Urbinati, C. R., Teng-Umnuay, P., Stenberg, M. G., Byrne, B. J., Thornton, C. A., and Swanson, M. S. Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J*, 19(17):4439–4448, 2000.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. Network motifs: simple building blocks of complex networks. *Sci STKE*, 298(5594):824, 2002.

Mitchell, P. J. and Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245(4916):371–378, 1989.

Mitsui, K., Nakayama, H., Akagi, T., Nekooki, M., Ohtawa, K., Takio, K., Hashikawa, T., and Nukina, N. Purification of polyglutamine aggregates and identification of elongation factor-1α and heat shock protein 84 as aggregate-interacting proteins. *J Neurosci*, 22(21):9267–9277, 2002.

Montojo, J., Zuberi, K., Rodriguez, H., Kazi, F., Wright, G., Donaldson, S. L., Morris, Q., and Bader, G. D. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*, 26(22):2927–2928, 2010.

Mrowka, R., Patzak, A., and Herzel, H. Is there a bias in proteome research? *Genome Res*, 11(12):1971–1973, 2001.

Nakayama, M., Kikuno, R., and Ohara, O. Protein-protein interactions between large proteins: two-hybrid screening using a functionally classified library composed of long cDNAs. *Genome Res*, 12(11):1773–1784, 2002.

Nepusz, T., Yu, H., and Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods*, 9(5):471–472, 2012.

Nucifora, F. C., Sasaki, M., Peters, M. F., Huang, H., Cooper, J. K., Yamada, M., Takahashi, H., Tsuji, S., Troncoso, J., Dawson, V. L., Dawson, T. M., and Ross, C. A. Interference by huntingtin and atrophin-1 with cbp-mediated transcription leading to cellular toxicity. *Sci STKE*, 291(5512):2423, 2001.

Nye, T. M. W., Berzuini, C., Gilks, W. R., Babu, M. M., and Teichmann, S. A. Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, 21(7):993–1001, 2005.

Oliver, S. Guilt-by-association goes global. *Nature*, 403(6770):601–603, 2000.

Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., Gnad, F., Cox, J., Jensen, T. S., Nigg, E. A., Brunak, S., and Mann, M. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci STKE*, 3(104):ra3, 2010.

Oppermann, F. S., Gnad, F., Olsen, J. V., Hornberger, R., Greff, Z., Kéri, G., Mann, M., and Daub, H. Large-scale proteomics analysis of the human kinome. *Mol Cell Proteomics*, 8(7):1751–1764, 2009.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H. W., Ruepp, A., and Frishman, D. The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21(6): 832–834, 2005.

Pal, S., Santos, A., Rosas, J. M., Ortiz-Guzman, J., and Rosas-Acosta, G. Influenza A virus interacts extensively with the cellular SUMOylation system during infection. *Virus Res*, 158(1):12–27, 2011.

Pastor-Satorras, R., Smith, E., and Solé, R. V. Evolving protein interaction networks through gene duplication. *J Theor Biol*, 222(2):199–210, 2003.

Pauli, E. K., Schmolke, M., Wolff, T., Viemann, D., Roth, J., Bode, J. G., and Ludwig, S. Influenza A virus inhibits type I IFN signaling via NF-κB-dependent induction of SOCS-3 expression. *PLoS Pathog*, 4(11):e1000196, 2008.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371, 2003.

Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A., and Cesareni, G. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, 6 Suppl 4:S21, 2005.

Perutz, M. F. Glutamine repeats and inherited neurodegenerative diseases: molecular aspects. *Curr Opin Struct Biol*, 6(6):848–858, 1996.

Perutz, M. F., Staden, R., Moens, L., and De Baere, I. Polar zippers. *Curr Biol*, 3(5): 249, 1993.

Petrakis, S., Raskó, T., Russ, J., Friedrich, R. P., Stroedicke, M., Riechers, S. P., Muehlenberg, K., Möller, A., Reinhardt, A., Vinayagam, A., Schaefer, M. H., Boutros, M., Tricoire, H., Andrade-Navarro, M. A., and Wanker, E. E. Identification of Human Proteins That Modify Misfolding and Proteotoxicity of Pathogenic Ataxin-1. *PLoS Genet*, 8(8):e1002897, 2012.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229, 2001.

Qi, Y., Bar-Joseph, Z., and Klein-Seetharaman, J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Struct , Funct , Bioinf*, 63(3):490–500, 2006.

Rachlin, J., Cohen, D. D., Cantor, C., and Kasif, S. Biological context networks: a mosaic view of the interactome. *Mol Syst Biol*, 2:66, 2006.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

Reddy, K., Tam, M., Bowater, R. P., Barber, M., Tomlinson, M., Nichol Edamura, K., Wang, Y. H., and Pearson, C. E. Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. *Nucleic Acids Res*, 39(5):1749–1762, 2011.

Reiner, A., Albin, R. L., Anderson, K. D., D'Amato, C. J., Penney, J. B., and Young, A. B. Differential loss of striatal projection neurons in Huntington disease. *Proc Natl Acad Sci U S A*, 85(15):5733, 1988.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–1032, 1999.

Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. Inferring protein domain interactions from databases of interacting proteins. *Genome Biol*, 6(10):R89, 2005.

Rives, A. W. and Galitski, T. Modular organization of cellular networks. *Proc Natl Acad Sci U S A*, 100(3):1128, 2003.

Ross, C. A. Intranuclear neuronal inclusions: a common pathogenic mechanism for glutamine-repeat neurodegenerative diseases? *Neuron*, 19(6):1147–1150, 1997.

Ross, E. D., Minton, A., and Wickner, R. B. Prion domains: sequences, structures and interactions. *Nat Cell Biol*, 7(11):1039–1044, 2005.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J., Guo, Y., Heriche, J. K., Hu, Y., Kristiansen, K., Li, R., Liu, T., Moses, A., Qin, J., Vang, S., Vilella, A. J., Ureta-Vidal, A., Bolund, L., Wang, J., and Durbin, R. TreeFam: 2008 Update. *Nucleic Acids Res*, 36(Database issue):D735–40, 2008.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O. N., Stumpflen, V., and Mewes, H. W. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, 36(Database issue):D646–50, 2008.

Said, M. R., Begley, T. J., Oppenheim, A. V., Lauffenburger, D. A., and Samson, L. D. Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 101(52):18006, 2004.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–51, 2004.

Schaefer, M. H., Fontaine, J. F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. HIPPIE: integrating protein interaction networks with experiment based quality scores. *PloS One*, 7(2):e31826, 2012a.

Schaefer, M. H., Wanker, E. E., and Andrade-Navarro, M. A. Evolution and function of CAG/polyglutamine repeats in protein–protein interaction networks. *Nucleic Acids Res*, 40(10):4273–4287, 2012b.

Schaefer, M. H., Lopes, T. J., Mah, N., Shoemaker, J. E., Matsuoka, Y., Fontaine, J. F., Louis-Jeune, C., Eisfeld, A. J., Neumann, G., Perez-Iratxeta, C., Kawaoka, Y., Kitano, H., and Andrade-Navarro, M. A. Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput Biol*, 9(1):e1002860, 2013.

Schmitz, N., Kurrer, M., Bachmann, M. F., and Kopf, M. Interleukin-1 is responsible for acute lung immunopathology but increases survival of respiratory influenza virus infection. *J Virol*, 79(10):6441–6448, 2005.

Scott, M. S., Perkins, T., Bunnell, S., Pepin, F., Thomas, D. Y., and Hallett, M. Identifying regulatory subnetworks for a set of genes. *Mol Cell Proteomics*, 4(5):683–692, 2005.

Seal, R. L., Gordon, S. M., Lush, M. J., Wright, M. W., and Bruford, E. A. genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*, 39(Database issue): D514–9, 2011.

Seet, B. T., Dikic, I., Zhou, M. M., and Pawson, T. Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol*, 7(7):473–483, 2006.

Sen, S., Dash, D., Pasha, S., and Brahmachari, S. K. Role of histidine interruption in mitigating the pathological effects of long polyglutamine stretches in SCA1: A molecular approach. *Protein Sci*, 12(5):953–962, 2003.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, 2003.

Shapira, S. D., Gat-Viks, I., Shum, B. O. V., Dricot, A., De Grace, M. M., Wu, L., Gupta, P. B., Hao, T., Silver, S. J., Root, D. E., Hill, D. E., Regev, A., and Hacohen, N. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*, 139(7):1255–1267, 2009.

Shy, M. E., Jáni, A., Krajewski, K., Grandis, M., Lewis, R. A., Li, J., Shy, R. R., Balsamo, J., Lilien, J., Garbern, J. Y., and Kamholz, J. Phenotypic clustering in MPZ mutations. *Brain*, 127(2):371–384, 2004.

Simon, M. and Hancock, J. M. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol*, 10(6):R59, 2009.

Stark, C., Breitkreutz, B. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K., and Tyers, M. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*, 39(Database issue):D698–704, 2011.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

Stenmark, H., Aasland, R., and Driscoll, P. C. The phosphatidylinositol 3-phosphate-binding FYVE finger. *FEBS Lett*, 513(1):77–84, 2002.

Strand, M., Prolla, T. A., Liskay, R. M., and Petes, T. D. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, 365(6443):274–276, 1993.

Suhr, S. T., Senut, M. C., Whitelegge, J. P., Faull, K. F., Cuizon, D. B., and Gage, F. H. Identities of sequestered proteins in aggregates from cells with induced polyglutamine expression. *J Cell Biol*, 153(2):283–294, 2001a.

Suhr, S. T., Senut, M. C., Whitelegge, J. P., Faull, K. F., Cuizon, D. B., and Gage, F. H. Identities of sequestered proteins in aggregates from cells with induced polyglutamine expression. *J Cell Biol*, 153(2):283–294, 2001b.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, 39(Database issue):D561–8, 2011.

Szretter, K. J., Gangappa, S., Lu, X., Smith, C., Shieh, W. J., Zaki, S. R., Sambhara, S., Tumpey, T. M., and Katz, J. M. Role of host cytokine responses in the pathogenesis of avian H5N1 influenza viruses in mice. *J Virol*, 81(6):2736–2744, 2007.

Ta, H. X. and Holm, L. Evaluation of different domain-based methods in protein interaction prediction. *Biochem Biophys Res Commun*, 390(3):357–362, 2009.

Tang, T. S., Slow, E., Lupu, V., Stavrovskaya, I. G., Sugimori, M., Llinas, R., Kristal, B. S., Hayden, M. R., and Bezprozvanny, I. Disturbed Ca2+ signaling and apoptosis of medium spiny neurons in Huntington's disease. *Proc Natl Acad Sci U S A*, 102(7): 2602–2607, 2005.

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. L. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol*, 27(2):199–204, 2009.

Tran, P. B. and Miller, R. J. Aggregates in neurodegenerative disease: crowds and power? *Trends Neurosci*, 22(5):194–197, 1999.

Truant, R., Atwal, R. S., and Burtnik, A. Nucleocytoplasmic trafficking and transcription effects of huntingtin in Huntington's disease. *Prog Neurobiol*, 83(4):211–227, 2007.

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., and Wodak, S. J. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)*, 2010: baq023, 2010.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

Venkatesan, K., Rual, J. F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K. I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A. S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabasi, A. L., and Vidal, M. An empirical framework for binary interactome mapping. *Nat Methods*, 6(1):83–90, 2009.

Vermeulen, M. and Selbach, M. Quantitative proteomics: a tool to assess cell differentiation. *Curr Opin Cell Biol*, 21(6):761–766, 2009.

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., Assmus, H. E., Andrade-Navarro, M. A., and Wanker, E. E. A Directed Protein Interaction Network for Investigating Intracellular Signal Transduction. *Sci STKE*, 4(189):rs8, 2011.

Vlasblom, J., Wu, S., Pu, S., Superina, M., Liu, G., Orsi, C., and Wodak, S. J. GenePro: a cytoscape plug-in for advanced visualization and analysis of interaction networks. *Bioinformatics*, 22(17):2178–2179, 2006.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.

von Mikecz, A. PolyQ fibrillation in the cell nucleus: who's bad? *Trends Cell Biol*, 19 (12):685–691, 2009.

Wachi, S., Yoneda, K., and Wu, R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 21 (23):4205–4208, 2005.

Waelter, S., Boeddrich, A., Lurz, R., Scherzinger, E., Lueder, G., Lehrach, H., and Wanker, E. E. Accumulation of mutant huntingtin fragments in aggresome-like inclusion bodies as a result of insufficient protein degradation. *Mol Biol Cell*, 12(5): 1393–1407, 2001.

Wang, H., Segal, E., Ben-Hur, A., Li, Q. R., Vidal, M., and Koller, D. InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biol*, 8(9):R192, 2007.

Wang, J., Huo, K., Ma, L., Tang, L., Li, D., Huang, X., Yuan, Y., Li, C., Wang, W., Guan, W., Chen, H., Jin, C., Wei, J., Zhang, W., Yang, Y., Liu, Q., Zhou, Y., Zhang, C., Wu, Z., Xu, W., Zhang, Y., Liu, T., Yu, D., Zhang, Y., Chen, L., Zhu, D., Zhong, X., Kang, L., Gan, X., Yu, X., Ma, Q., Yan, J., Zhou, L., Liu, Z., Zhu, Y., Zhou, T., He, F., and Yang, X. Toward an understanding of the protein interaction network of the human liver. *Mol Syst Biol*, 7:536, 2011.

Wang, X. and Huang, L. Identifying dynamic interactors of protein complexes by quantitative mass spectrometry. *Mol Cell Proteomics*, 7(1):46–57, 2008.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., and Morris, Q. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*, 38(suppl 2):W214–W220, 2010.

Warrick, J. M., Paulson, H. L., Gray-Board, G. L., Bui, Q. T., Fischbeck, K. H., Pittman, R. N., and Bonini, N. M. Expanded polyglutamine protein forms nuclear inclusions and causes neural degeneration in Drosophila. *Cell*, 93(6):939–949, 1998.

Wass, M. N., Fuentes, G., Pons, C., Pazos, F., and Valencia, A. Towards the prediction of protein interaction partners using physical docking. *Mol Syst Biol*, 7:469, 2011.

Weatheritt, R. J., Davey, N. E., and Gibson, T. J. Linear motifs confer functional diversity onto splice variants. *Nucleic Acids Res*, 40(15):7123–7131, 2012.

Wehr, M., Reinecke, L., Botvinnik, A., and Rossner, M. Analysis of transient phosphorylation-dependent protein-protein interactions in living mammalian cells using split-TEV. *BMC Biotechnol*, 8(1):55, 2008.

Wells, J. A. and McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*, 450(7172):1001–1009, 2007.

Wen-hsien, L., Wei-chung, L., and Ming-jing, H. Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein-protein interaction networks. *BMC Systems Biology*, 3(1):32, 2009.

Whan, V., Hobbs, M., McWilliam, S., Lynn, D., Lutzow, Y., Khatkar, M., Barendse, W., Raadsma, H., and Tellam, R. Bovine proteins containing poly-glutamine repeats are often polymorphic and enriched for components of transcriptional regulatory complexes. *BMC Genomics*, 11(1):654, 2010.

Williams, A. D., Portelius, E., Kheterpal, I., Guo, J., Cook, K. D., Xu, Y., and Wetzel, R. Mapping Aβ amyloid fibril secondary structure using scanning proline mutagenesis. *J Mol Biol*, 335(3):833–842, 2004.

Wood, S. J., Wetzel, R., Martin, J. D., and Hurle, M. R. Prolines and Aamyloidogenicity in Fragments of the Alzheimer's Peptide. beta./A4. *Biochemistry (Mosc)*, 34(3):724–730, 1995.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C. L., Haase, J., Janes, J., Huss, J. W., and Su, A. I. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol*, 10(11):R130, 2009.

Xing, Z., Harper, R., Anunciacion, J., Yang, Z., Gao, W., Qu, B., Guan, Y., and Cardona, C. J. Host immune and apoptotic responses to avian influenza virus H9N2 in human tracheobronchial epithelial cells. *Am J Respir Cell Mol Biol*, 44(1):24, 2011.

Yang, L., Walker, J. R., Hogenesch, J. B., and Thomas, R. S. NetAtlas: A Cytoscape plugin to examine signaling networks based on tissue gene expression. *In Silico Biol*, 8(1):47–52, 2008.

Yang, Z. R., Thomson, R., McNeil, P., and Esnouf, R. M. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, 21(16):3369–3376, 2005.

Yook, S. H., Oltvai, Z. N., and Barabási, A. L. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

Yosef, N., Ungar, L., Zalckvar, E., Kimchi, A., Kupiec, M., Ruppin, E., and Sharan, R. Toward accurate reconstruction of functional protein networks. *Mol Syst Biol*, 5:248, 2009.

Yosef, N., Zalckvar, E., Rubinstein, A. D., Homilius, M., Atias, N., Vardi, L., Berman, I., Zur, H., Kimchi, A., Ruppin, E., and Sharan, R. ANAT: A Tool for Constructing and Analyzing Functional Protein Networks. *Sci STKE*, 4(196):pl1, 2011.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J. F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A. S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabasi, A. L., Tavernier, J., Hill, D. E., and Vidal, M. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

Yu, J., Guo, M., Needham, C. J., Huang, Y., Cai, L., and Westhead, D. R. Simple sequence-based kernels do not predict protein–protein interactions. *Bioinformatics*, 26(20):2610–2614, 2010.

Zhai, W., Jeong, H., Cui, L., Krainc, D., and Tjian, R. In vitro analysis of huntingtin-mediated transcriptional repression reveals multiple transcription factor targets. *Cell*, 123(7):1241–1253, 2005.

Zhao, L. and Chmielewski, J. Inhibiting protein-protein interactions using designed molecules. *Curr Opin Struct Biol*, 15(1):31–34, 2005.

List of Figures

List of Tables

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, März 2013 Martin Schaefer